Copyright and Other Legal Notices

You may print multiple copies of these materials for your own personal use, including use in your classes and/or sharing with individual colleagues as long as the author's name and institution, and a notice that the materials were obtained from the website of the Society for the Teaching of Psychology (STP) <http://teachpsych.org/> appear on the copied document. For research and archival purposes, public libraries and libraries at schools, colleges, universities and similar educational institutions may print and store in their research or lending collections multiple copies of this compendium as a whole without seeking further permission of STP (the authors would appreciate receiving a pro forma notice of any such library use). No other permission is granted to you to print, copy, reproduce, or distribute additional copies of these materials. Anyone who wishes to print, copy, reproduce, or distribute copies for other purposes must obtain the permission of the individual copyright owners. Particular care should be taken to seek permission from the respective copyright holder(s) for any commercial or "for profit" use of these materials.

Suggested Reference Format


Feedback

Feedback regarding the content of this book should be directed toward the editors. They are solely responsible for the substance of the text. Feedback regarding technical matters of formatting or accessibility of this text via the online environment of the Internet should be directed to the Internet Editor. If you have any complaints or difficulties in accessing these materials, be sure to provide as detailed a description of your problem(s) as you can; you should include information about the browser you are using (e.g., Firefox, Safari) and its version number well as the type of computer you are using and its operating system (e.g., Mac PowerBook 4 running MacOS 10.3.2).
TEACHING STATISTICS AND RESEARCH METHODS: TIPS FROM *TOP*

Edited by

Sherri L. Jackson
Jacksonville University

Richard A. Griggs
University of Florida
FOREWORD

This book follows in the footsteps of the first two volumes in the *Handbook for Teaching Statistics and Research Methods* series. The editors of those two volumes underscored the importance of data analysis and research methodology to empirical disciplines such as psychology. Our purpose in preparing this volume is the same as that of the previous editors—to assist those who teach courses whose contents contain statistics and research methodology. A current, readily accessible, and organized collection of *Teaching of Psychology (ToP)* articles on teaching the statistics and research methods courses should be an invaluable resource to psychology teachers, and indeed, this has proved to be the case for the first two volumes. We hope that the same holds true for this volume. To emphasize the resource nature of this new volume in the series, we abandoned the more formal previous titles and entitled this new volume, *Teaching Statistics and Research Methods: Tips from ToP*.

For continuity, we maintained the overall structure of the previous volumes by having two main sections (Statistics and Research Methods) and kept as many of the subsections for each main section as possible given the content of the current set of articles. Four of the previous Statistics subsections (Generating Data Sets, Illustrating Statistical Concepts, Examining Statistical Tests, and Developing Students’ Skills) remain, but there are three new ones (Surveying Statistical Training in Psychology, Implementing Teaching Strategies, and Using Technology in Teaching Statistics). Given the increasing role of computers and other technologies in teaching, it was necessary for articles concerned with these technologies to become a separate subsection in both main sections.

Four of the 10 subsections in the Research Methods section (Implementing Teaching Strategies, Demonstrating Systematic Observation and Research Design, Fostering Students’ Research and Presentations, and Teaching Ethics) were maintained and another (Teaching Writing and Critical Thinking) was divided into two subsections, Teaching Writing and APA Style and Teaching Critical Thinking), because of the large number of articles on these two topics. Four new subsections were added—Analyzing Research Methods Textbooks, Using Technology in Teaching Research Methods, Assessing Training in Research Methods, and Assessing Students’ Attitudes Toward Scientific Thinking.

We provide a Brief Table of Contents so that you can more easily view the entire organizational structure of the book—the two major sections with all of the subsections for each one listed. The full Table of Contents includes a listing of the articles within each subsection as well as the citation information for each paper. The articles are ordered chronologically within each subsection.
according to their publication dates starting with the earliest one. The book is structured to maximize flexibility in its use. You can go directly to the topic(s) and article(s) of your choice. Thus, for example, if you are interested in using computer software in teaching your statistics course, you can go directly to the Using Technology in Teaching Statistics subsection and to any article within that subsection that you wish. We recommend that you first peruse the Brief Table of Contents to see the overall structure of the book and then visit the full Table of Contents to find what articles are included in each subsection of your interest.

The second volume in this series, published in 1999, included 92 articles published mainly in the period 1987 through 1998; five articles from 1985 and 1986 that had been overlooked in the first volume were also included. The article selection period for this volume thus ranges from 1999 through June 2012 (13½ years), a slightly longer publication period than that for the second volume. We conducted a thorough examination of all 53 regular issues of ToP during this period to identify articles appropriate for inclusion in this volume. In addition, we used ToP’s yearly Annotated Bibliographies on the Teaching of Psychology from 1999 through the final one in 2008 to verify that we had not omitted any relevant articles. Statistics and Research Methods and Research-Related Issues were categories in these bibliographies. This process resulted in the inclusion of 107 articles in this volume—50 on statistics and 57 on research methods.

Published in 1999, the second volume in this series celebrated ToP’s silver anniversary and the outstanding work of ToP’s Founding Editor, Robert S. Daniel, who guided ToP through its critical first 10 years. This book celebrates ToP’s upcoming 40th anniversary and the invaluable contributions of ToP’s two Editors Emeriti, Charles L. Brewer (1985-1996) and Randolph A. Smith (1997-2008), who led ToP through adolescence and its adulthood years before turning the editorial reins over to Andrew N. Christopher in 2009 to guide it on into middle age. As with the previous two volumes, this book contains the work of many dedicated teachers. We would like to thank all of them for sharing their ideas and research with us. We would also like to thank Bill Buskist and Jeff Stowell for their invaluable help in bringing this project to fruition.

Sherri L. Jackson  
Jacksonville University

Richard A. Griggs  
University of Florida
**BRIEF TABLE OF CONTENTS**

**FRONT MATTER**

<table>
<thead>
<tr>
<th>Title Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreward</td>
</tr>
<tr>
<td>Brief Table of Contents</td>
</tr>
<tr>
<td>Table of Contents</td>
</tr>
</tbody>
</table>

**SECTION I. STATISTICS**

1. Generating Data Sets
2. Illustrating Statistical Concepts
3. Examining Statistical Tests
4. Developing Students’ Skills
5. Surveying Statistical Training in Psychology
6. Implementing Teaching Strategies
7. Using Technology in Teaching Statistics

**SECTION II. RESEARCH METHODS**

1. Analyzing Research Methods Textbooks
2. Implementing Teaching Strategies
3. Demonstrating Systematic Observation and Research Design
4. Teaching Critical Thinking
5. Teaching Writing and APA Style
6. Fostering Students’ Research and Presentations
7. Using Technology in Teaching Research Methods
8. Teaching Ethics
9. Assessing Training in Research Methods
10. Assessing Students’ Attitudes Toward Scientific Thinking
TABLE OF CONTENTS

Front Matter and Appendix

Title Page
Foreword
Brief Table of Contents
Full Table of Contents

Section I. Statistics

1. Generating Data Sets


2. Illustrating Statistical Concepts


3. Examining Statistical Tests


4. Developing Students’ Skills


5. Surveying Statistical Training in Psychology


6. Implementing Teaching Strategies


7. Using Technology in Teaching Statistics


**Section II: Research Methods**

1. **Analyzing Research Methods Textbooks**


2. **Implementing Teaching Strategies**


3. **Demonstrating Systematic Observation and Research Design**


4. **Teaching Critical Thinking**


5. Teaching Writing and APA Style


6. Fostering Students’ Research and Presentations


7. Using Technology in Teaching Research Methods


8. Teaching Ethics


9. Assessing Training in Research Methods


10. Assessing Students’ Attitudes Toward Scientific Thinking


AnoGen: A Program for Generating ANOVA Data Sets

Jeff Miller
University of Otago

In this article, I describe AnoGen, a free program for generating ANOVA data sets on IBM–PC compatible computers running the DOS operating system.

Rapid advances in computer technology have revolutionized the practice of data analysis in the last 20 years. Researchers may now use powerful statistical analysis packages available for personal computers, thus avoiding hand calculation.

Statistics teachers have traditionally faced a problem of data construction—as distinct from analysis—that is not so easily solved with such programs. With computers available for analysis, of course, teachers prefer to use large, real data sets rather than small, artificial ones (e.g., Thompson, 1994) in many circumstances. Yet some teachers may still want to construct small, simple data sets for a variety of purposes. Such data sets can be valuable not only for in-class computational examples but also for examinations, homework assignments, and even supplementary practice for students having difficulty mastering the techniques. If the data set is to be examined within a realistic problem context, it must contain realistic numbers (e.g., IQ scores in the range of 70 to 130) and it needs to have particular means, variances, and effect sizes appropriate to a given example study. As Read and Riley (1983) noted, it is often also desirable to ease rounding difficulties by arranging for means to be whole numbers or to be exact to one or two decimal places because simpler calculations are easier to perform and thus distract less from the underlying concepts. Construction of data sets satisfying several such constraints can be a tedious process, generally proceeding by trial and error, at best aided by a few simple numerical tricks.

Computers can ease the task of data construction just as they ease its analysis (e.g., Beins, 1989; McGown & Spencer, 1980). I wrote a computer program, called AnoGen, to automate the process of generating data sets for ANOVA designs. The program operates within DOS and Windows compatible computers, and it is freely available for noncommercial use via the Internet.

AnoGen can be run in either of two modes: student or teacher. In either mode, the user first specifies the desired number of between- and within-subjects factors, the number of levels of each factor, and the number of participants. In student mode, the computer then generates the data. The student can view this data set or save it to a file, along with its associated ANOVA table and various intermediate steps in its analysis (e.g., the ANOVA model, the cell and marginal means).
In teacher mode, the user interactively sets the means, effect sizes, and variability by assigning values to the parameters of the linear model appropriate to the specified design (or, in the case of random terms in the model, by assigning ranges of values). As the teacher adjusts the parameter values, the data set, cell and marginal means, and ANOVA table are constantly updated, and these may be inspected at any time to see how the desired constraints are being met.

The capabilities of AnoGen should be sufficient for generating any ANOVA data set that one might want to analyze by hand. The primary limitation is that the number of data values times the number of terms in the model cannot exceed 32,766. This limitation still allows quite large data sets, including, for example, three between-subjects factors, two within-subjects factors, and 80 participants. In addition, the generated data values are limited to the range of $\pm 32,766$ and may contain only five digits of precision.

My teaching assistants and I have used various versions of AnoGen for more than 15 years to generate hundreds of data sets. It adds freshness and enthusiasm to my lectures because I can translate a new idea for an interesting example (e.g., obtained from the newspaper, a seminar, or a journal article) into an appropriate data set, with a complete analysis, usually within 5 min. Equally important, I can generate computational problems for homework and exams quickly, and the program generates numbers that create minimal opportunities for rounding error. In addition to the data set, the program outputs the correct ANOVA summary table and, optionally, the correct form of the general linear model, the estimation equations needed to estimate the terms in the model, and a decomposition matrix showing how each data value is partitioned into its numerical components. I made AnoGen directly available to students for the first time in my most recent ANOVA class, and approximately 10% of the students used it to generate practice problems on their own. No difficulties with the user interface were reported. I hope that other teachers and students of ANOVA will also find the program useful.

References


Notes

1. The AnoGen program is available from common FTP sites, including simtel and garbo. The current version is 1.3, available in the file AnoGen13.zip (e.g., http://www.simtel.net/pub/simtelnet/msdos/statscs/Anogen13.zip). It comes with a detailed manual for teacher use and a simplified manual for distribution to students, each in postscript, hypertext markup language, and public document format.

2. Send correspondence to Jeff Miller, Department of Psychology, University of Otago, Dunedin, New Zealand; e-mail: miller@otago.ac.nz.
The Statistically Marvelous Medical Growth Chart: A Tool for Teaching Variability

Kenneth A. Weaver
Emporia State University

Plotting girls' and boys' weights on a medical growth chart in the introductory statistics course illustrates variability, the normal distribution, percentiles, z scores, outliers, bivariate graphing, and simple regression. The chart presents the spread of weights for newborns through 36 months, includes percentile scores, and represents a bivariate distribution with age on the abscissa and weight on the ordinate. Students plot their own weights to understand how the chart works and then plot the weights of a selected boy and girl to understand how the chart identifies outliers for follow-up tests on hormone levels, nutrition, and intellectual development. Instructors in other psychology courses (e.g., developmental, child, abnormal, introductory, and educational psychology) may also find the chart useful when covering infant development.

Medical growth charts, published by pharmaceutical companies such as Mead Johnson & Company, can be used to demonstrate a variety of statistical concepts in psychology. The charts are based on data collected from 1963 through 1975 by the National Center for Health Statistics (NCHS) using probability samples including over 20,000 children (Hamill et al., 1979). The NCHS is currently revising the growth charts based on data collected between 1988 and 1994 from 33,994 persons in a nationwide probability sample during the third National Health and Nutrition Examination Survey (NCHS, 1998; Springen, 1998).

The growth charts are commonly available as a packet from pediatricians and local health departments. Figure 1 is the weight chart for boys from birth through 36 months; the girls' weight chart is virtually identical.

Teaching Activities Using the Charts

Illustrating the Spread of Weights

The week before this activity, I ask students to determine their birth weights. The day of the activity, I ask students to estimate the range of weights for a sample of newborns. I provide the following prompts to help: "For a group of healthy newborns, what are some of the possible lowest weights?" and "For a group of healthy newborns, what are some of the possible highest weights?" I also ask students to estimate the equivalent range of weights for the 36-month-old children. I then ask the class which range of weights is larger: the distribution of weights for the newborns or the distribution of weights for the toddlers.

Almost all students answer without hesitation that the range of weights for 36-month-old children is greater. The visual referent of the greater spread of weights with age appears to confirm the students' expectation.

I point out that the absolute range of weights for 36-month-old children is greater than newborns. However, the 95th relative to the 5th percentile weights for newborns and 36-month-old children are 1.63 and 1.40 times larger, respectively (based on Table 2 in Hamill et al., 1979). Relatively speaking, newborns' weights are more variable.

Connecting Percentiles and z Scores to Spread

I display a transparency of one of the weight charts. With an overhead transparency pen, I draw a normal curve straddling the percentiles and ask three orienting questions: What percentage of the weights are located above the 50th percentile, what does an infant's weight located at the 75th percentile mean, and why are the widths between the 10th and 25th percentiles and between the 25th and 50th percentiles the same when the differences between these two pairs of numbers (15% and 25%, respectively) are unequal?

Students plot their birth weights on the gender-appropriate chart. After plotting their birth weights, they use the right ordinate of Figure 1 to convert their birth weights to percentiles and then hypothesize the sign and size of the corresponding z score. Students check their answers using a table for converting percentiles to z scores from their textbook (Table B in Sprinthall, 1997, p. 524).
Variability as a Tool for Health

After asking students the medical purpose of the chart, I show a transparency of a 1996 article (available at http://www.thonline.com/th/news/1996/th0813/stories/21732.htm) about a male toddler who weighed 68 pounds at 17 months. Students locate the toddler’s weight on their copy of the boys’ chart, and I ask how frequently the toddler’s weight occurs relative to the chart weights of the 17-month-old children. At the 99th percentile, the toddler’s weight is an extreme outlier having a very low frequency of occurrence. Extreme variability of this sort often signals a possible problem that may lead a pediatrician to order follow-up tests on growth hormones, nutrition, or intellectual development. The article noted that in the case at hand an endocrinologist reported nothing abnormal.

Students also locate the 2 lb 5 oz birth weight of the lightest McCaughey septuplet (Schindehette, Fowler, Grisby, & Breu, 1997) on their copy of the girls’ chart. In contrast to the toddler, the septuplet’s low weight corresponds to a low rather than high percentile score. I again display the transparency of the weight chart and direct students’ attention to the normal curve drawn earlier. Students then connect low frequency of occurrence with both very low and very high percentile scores in the tails of the normal distribution.

I ask for adjectives to describe these two children’s weights, and students respond with terms such as rare, extreme, strange, and very unusual. Their responses show that the students have grasped the statistical underpinnings of these words.

Introducing Bivariate Graphing

With the chart still displayed, I point out the X axis and Y axis. I present to the students $Y = f(X)$ and ask why weight is on the Y axis and not the X axis. The axes are so labeled because the values of weight are a function of the child’s age and thus depend on age. Linking the term function with the term depend identifies the variable graphed on the Y axis as the dependent variable. This content foreshadows later coverage of the dependent variable for the t test, analysis of variance, and correlation. I ask why weight is given in pounds on the left ordinate and kilograms on the right ordinate. Health officers in other nations can use the chart to monitor infant development.

Other Charts in the Packet

The packet includes growth charts for length and head circumference by age and a graph plotting weight by length. For simple regression, the graph clearly illustrates the linear relationship between weight and length.

Variability Applied to Psychology

The final activity involves showing the class a transparency of an anonymous score report from the Comprehensive Test of Basic Skills (1990), which I obtained from the local middle school counselor. The report labels sections of the normal distribution as below average, average, and above average. As with the growth chart, I draw the normal distribution on the transparency straddling the percentiles. Students see that student achievement scores, like children’s weights, are interpreted as degrees of variability from the mean. Like pediatricians, educators often take a score’s extreme variability to indicate a possible need for remediation or enrichment (Slavin, 1997).

Student Evaluations

Students ($N = 108$) in three classes evaluated the activity’s purpose, presentation, and interest level using a grading scale ranging from 0 (F) to 100 (A+). Students gave mean responses of 92 ($SD = 6$), 93 ($SD = 6$), and 93 ($SD = 7$), respectively.

Conclusions

Giesbrecht, Sell, Scialfa, Sandals, and Ehlers (1997) included measures of center, measures of variability, the normal distribution, percentile scores, bivariate plotting techniques, outliers, and simple regression as essential topics to cover in the introductory statistics course. The instructor can elaborate all of these topics using medical growth charts.

Instructors in courses other than statistics may also find the charts useful. Child psychology and developmental psychology courses cover fetal and infant growth and often address the physical and cognitive problems of very low birth weight, low birth weight, and small-for-gestational-age infants (e.g., see Feldman, 1997, pp. 92–93). The abnormal psychology course may also refer to the association between low birth weight and schizophrenia (Kendall & Hammen, 1998). Low birth weight is also a possible topic in the developmental sections of introductory and educational psychology courses. The charts would complement the coverage by showing students the extreme variability of infants’ low birth weights, the rarity of their occurrence, and the diagnostic value of the charts for physical and cognitive problems.

References

National Center for Health Statistics. (1998). Third National Health and Nutrition Examination Survey (NHANES III) [Pub-
In this article, I describe a course to prepare undergraduates for graduate study in psychology. The course focuses on activities that develop and refine specific skills required by the application process. Over half the students who replied to a follow-up survey regarding the course applied to graduate school; of these, over 90% were accepted. The course is well-received by students and faculty alike and appears effective in preparing students to make application to graduate school.

Most teachers of psychology, especially those teaching upper division courses, receive queries by students interested in applying to psychology graduate programs. Students are curious about specific aspects of this process: How to match their career interests and graduate programs that offer degrees in those areas; (c) become more familiar with the GRE; (d) prepare a vita; (e) draft a letter of intent; and (f) successfully complete a personal interview. The basic components of the course parallel these objectives; a brief description of these activities appears in Table 1.

The graded activities involve two written assignments that students may turn in as many times as necessary to earn an A on each. The original draft and subsequent revisions of these assignments receive written feedback, and students meet with me to review particular problems they have in completing assignments. The nongraded activities are aimed at prompting students to get started with the application process and preparing them to provide a rationale for why they wish to attend graduate school. Although ungraded, students take these activities seriously as indicated by the promptness and detail with which they complete the activities. In general, students correctly judge these activities as being central to completing the application process successfully.

### Course Objectives and Activities

The seminar attempts to help students (a) become more knowledgeable about all aspects of graduate school, from starting the application process through completing the doctorate; (b) identify a specialty area within psychology that matches their career interests and graduate programs that offer degrees in those areas; (c) become more familiar with the GRE; (d) prepare a vita; (e) draft a letter of intent; and (f) successfully complete a personal interview. The basic components of the course parallel these objectives; a brief description of these activities appears in Table 1.

The graded activities involve two written assignments that students may turn in as many times as necessary to earn an A on each. The original draft and subsequent revisions of these assignments receive written feedback, and students meet with me to review particular problems they have in completing assignments. The nongraded activities are aimed at prompting students to get started with the application process and preparing them to provide a rationale for why they wish to attend graduate school. Although ungraded, students take these activities seriously as indicated by the promptness and detail with which they complete the activities. In general, students correctly judge these activities as being central to completing the application process successfully.

<table>
<thead>
<tr>
<th>Course Activity</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vita*</td>
<td>Documents relevant biographical, educational, research, professional, and reference information.</td>
</tr>
<tr>
<td>Letter of intent*</td>
<td>One-page, single spaced; most difficult of all assignments; constructive feedback is useful to students in drafting solid letter.</td>
</tr>
<tr>
<td>Specialty area*</td>
<td>A written description of the subfield of psychology to be studied and justification for choosing that area.</td>
</tr>
<tr>
<td>Select graduate programs*</td>
<td>A list of each student’s top 10 schools that offer programs in the specialty area.</td>
</tr>
<tr>
<td>Completed application*</td>
<td>A completed application packet (minus reference letters) to each student’s top choice school.</td>
</tr>
<tr>
<td>Mock interview*</td>
<td>A 10 to 15 min interview consisting of 6 to 10 questions that might be asked during an actual interview.</td>
</tr>
</tbody>
</table>

*aGraded. *Nongraded.
I require a second form of writing assignment called the “one-page press release” (Beins, 1993). Periodically, I ask students to read a preselected refereed journal article. The article includes statistical analyses currently being taught in the class. I ask students to summarize the article into a one-page press release using no statistical terminology. The assignment helps to develop students’ reading and writing skills as well as their understanding of professional literature.

Another form of writing assignment consists of a learning assessment journal (Qualters & Dolinsky, 1995). The journal assignment asks students to monitor their own learning processes. About every 2 weeks, I assign students a general topic to write about. Topics include their feelings toward statistics, experience with computers, strategies to solve problems, and studying strategies. The journal entries typically range in length from 100 to 250 words. The entries are supposed to be evaluative and demonstrate perceptions of learning in the course. I read and comment on their entries. The purpose of the journal is to allow students to develop self-knowledge of their learning, not only in the statistics course but other courses as well. From my perspective, the journal is also helpful in monitoring student mastery of topics. It allows me to advise students on effective learning strategies and to act as a motivator for students who admit to having frustrations and difficulties.

Conclusions

My goal in creating an active learning environment is to have students become more involved in their learning and develop their critical thinking skills. Is this a better method? Student grades still fit a normal distribution and teaching evaluations remain some of the highest in the college. Student comments, however, are considerably different. Students speak of understanding the material rather than just memorizing facts. They describe feelings of pride and self-confidence at being able to independently solve problems. They also describe insights regarding strengths and weaknesses in their learning styles. I firmly believe these learning achievements were the result of using an active learning approach to teaching statistics.

References


Notes

1. An earlier report of this article was presented at the American Psychological Association meeting, San Francisco, August 1998.
2. Send correspondence to Beverly Dolinsky, Division of Arts and Sciences, Endicott College, 376 Hale Street, Beverly, MA 01915; e-mail: bdolinsk@endicott.edu.

Statistically Lively Uses for Obituaries

Betsy Levonian Morgan
University of Wisconsin–La Crosse

I describe the benefits of using a real data set comprised of local obituaries to illustrate a variety of research issues, such as missing data, outliers, comparing means, and hypothesis testing. The inherent flaws in this type of data set give students hands-on experience with the concerns encountered in analysis of actual data.

Illustrating common principles with hands-on activities (e.g., Weaver, 1992; Zerbolio, 1989) and generalizing underlying principles to new situations (Evans, 1976) are two effective pedagogical approaches to promote the learning of statistical concepts. The activity described in this article combines both approaches by having students analyze a data set they constructed from information contained in obituaries. This activity also addresses several areas of concern to faculty who teach statistics, such as illustrating statistical concepts and developing students’ skills while reducing their fears (Ware & Brewer, 1999). By collecting obituaries over several days, I created a set of real data that students could easily enter into a computer program, analyze, and interpret.

I have used this activity with psychology undergraduates designing independent research projects and with secondary education math teachers as part of a continuing education program. All of these students had completed a course in statistics, but most were new to data sets and statistical software. Students’ summary appraisals of the activity were highly positive. Obituaries are rich in detail and represent a wide range of variability. The human-interest factors inherent in obituaries make them salient to students.

The Data Set

Students received a set of obituaries that I photocopied from a local newspaper. They assigned an identification num-
ber to each entry and then tried to identify and code the gender, age at death (in years), and the deceased’s number of children. Instructors who wish to use this activity should find approximately 50 obituaries, including one for an infant and some incomplete entries (e.g., no age at death or gender). The resulting data set comes rife with selection problems to solve. After resolving these problems and entering the data into SPSS for Windows 8.0 (1997) statistical software, the students answered a series of research questions requiring data analysis. Data entry and analyses took approximately 2 hr and occurred in an instructional computer laboratory.

Using obituaries may upset students. I forewarned students about the activity and gave them the dates of the newspapers in advance so that I could delete entries that a student wished to have removed. However, an instructor could use obituaries from another city or another time period or could change to an alternative data source such as wedding announcements or interest rates.

Student Learning Outcomes

1. Statistical software is user friendly. Many of my students showed traditional computer hesitancy, but the small sample size and small number of variables simplified data entry and made the entire process hands on. For data analyses, I posed a series of questions (e.g., “What is the average age at death for the women in this sample?”) rather than providing step-by-step instructions. The handouts had the questions on one side with detailed instructions on the other side for students who had difficulty.

2. Real data are messy. Obituaries give incomplete information. Students realized that they had to take into account the date of the newspaper to calculate the age at death if it was not given. There were duplicate data if a short obituary ran on one day and a fuller one ran the next day. Some obituaries did not include the information necessary to compute an age at death. All of these problems were frustrating to my students. Furthermore, obituaries that did not specifically state gender (by use of pronoun) and included a gender-neutral first name (e.g., Jan) forced students to make decisions. The students individually decided how to handle the gender issue. Most students chose to make educated guesses, but a few coded the variable as missing. Inevitably, these two groups produced different results from their analyses.

3. Outliers matter. The inclusion of an infant’s obituary provided two strong lessons. First, students had to decide how to code the infant’s age at death because it was unclear if it was a matter of days or months (we had an obituary that merely stated “the infant daughter of”). After we discussed the issues, students arbitrarily entered the infant’s age at death as 1 year. Second, the infant’s age skewed the age distribution for this sample. Students calculated the mean and standard deviation for the age at death, with and without the infant, to see how extreme numbers can influence these statistics.

Students also calculated the mean and standard deviation of age at death on one half of the obituaries (including the infant) and then repeated the analysis for the whole sample. These results illustrated how a smaller sample is more affected by extreme scores than a larger sample, leading to a discussion of sampling distributions. For hypothesis testing, we dropped the infant as an outlier using the quick rule of thumb of 3 SDs below or above the mean (e.g., Devore & Peck, 1997). The data set also had a young person (20 years old) who affected the analyses, but was not a classic outlier, that provided a nice lesson in variance.

4. Real questions are answered. The central analysis for this data set was a gender comparison on age at death. Unlike many research questions, there are population data available on this topic that allowed us to hypothesize that women should be significantly older at death than men. Instructors can discuss null and alternative hypotheses. Students found means and standard deviations for the total sample. They also learned how to use the software to separate the two groups and find the means for women and men separately. Finally, they performed a t test for significance testing and ran a one-way ANOVA to illustrate the relation between t tests and ANOVA. These data also could be used to review one-tailed versus two-tailed tests.

5. Correlations can be meaningful or nonsensical. The adage “correlation is not causation” was well illustrated by computing the correlation of the deceased’s number of children with age at death. Students enjoyed generating a variety of theories to support either a positive or negative relation. Number of children as a variable also provided another lesson in the problems with missing data or inaccurate data collection: It was unclear whether an individual in an obituary did not have children or whether children were just not mentioned. Finally, students created a nonsensical correlation (e.g., between identification number and age at death) to help show that statistical procedures and the software that compute them are tools that must be used wisely.

6. Results are summarized in a variety of ways. As a follow-up project, the undergraduates used the data analyses to write a method section including the handling of missing data and a series of result statements. They wrote two results sections, one with a table and one without, to explore how information in tables should not be redundant with text. They consulted the American Psychological Association (1994) Publication Manual to correctly report statistical results and format their paper.

Conclusions

I encourage instructors who introduce students to data entry, statistical software, and basic statistical analyses to consider obituaries as a data set. I believe that the activity would work well in a variety of teaching situations involving statistical principles. Instructors could reduce, expand, or modify this activity for use in introductory psychology, research de-
An Informal Seminar to Prepare the Best Undergraduates for Doctoral Programs in Psychology

William J. Lammers
University of Central Arkansas

In this article, I describe an informal seminar to prepare the best psychology undergraduates for application to and success in graduate school, with particular emphasis on doctoral programs. Student feedback, both formal and informal, suggests that this seminar can be an effective component of an overall advising program.

There are many components to undergraduate advising (Ware, 1992) and a variety of ways that undergraduate students can learn about graduate school and subsequent careers. Psychology departments use different strategies to prepare their undergraduates for graduate school and employment. Some departments rely on the students themselves, on individual advisors, on meetings and workshops offered by Psi Chi chapters or Psychology Clubs (e.g., Satterfield & Abramson, 1998), on courses or seminars specifically designed to accomplish this task (Buckalew & Lewis, 1982; Buskist, 1999; Davis, 1988; Dodson, Chastain, & Landrum, 1996; Oles & Cooper, 1988; Ware, 1988), and many rely on a mix of strategies.

Rationale for the Informal Seminar

As noted, several authors have described the type of information that can be presented via informal meetings of a psychology club or via a formal course. A long-term, noncredit seminar for this purpose has not been described in the literature. My department chose this approach for several reasons. Our faculty knew that even our best students were not aware of the highly competitive nature of graduate admission or the factors that strengthen an application. Students with strong academic credentials and the motivation to pursue a doctoral degree needed specific information before their senior year and needed individualized attention and encouragement to achieve a higher level of involvement in the discipline.

Each method for imparting information about doctoral programs has advantages and disadvantages. Offering a formal course for credit provides much information with structure and incentive for students to learn. Possible shortcomings can include an inability to target the relatively small group of students who can succeed in doctoral study, lack of a long-term mentor, and resistance on some campuses to offering such a course for credit (for a brief discussion of the latter issue, see Ware, 1988). Offering informal sessions hosted by a psychology club provides important information requiring few department resources. However, these sessions are typically infrequent, offer no mentoring, draw few students, and do not produce peer support groups.

I wanted a small, selective, and informal group. Without a formal course, I could personally invite a small number of our best students to become part of a group with common academic ambitions; I could be flexible with meeting times and places; I would not worry about grades; and I could mentor the students for the remainder of their undergraduate years. I hoped that a long-term commitment by such a small group would lead to strong support and friendships among the students, individualized attention to each student’s interests, and a system of peer advising whereby upper level students could give advice to lower level students. I did not believe that these goals could be effectively obtained with other modes of advising described in the literature.

Format for the Informal Seminar

Student Selection

At the beginning of the academic year, I review a list of all psychology majors that shows students’ classification, American College Testing scores, high school grade point average (GPA), cumulative college GPA, and psychology GPA. Beginning with the best set of scores, I select the top 20 students and send them a letter of invitation describing the purpose of the program.

From among the 20 invitations, several students typically have no plans for pursuing a doctoral degree. Several others are interested but do not follow through by attending meetings. Each year, approximately 6 students, who are primarily freshmen and sophomores, join the 8 to 12 students who are continuing their involvement.
Measuring School Spirit: A National Teaching Exercise

The School Spirit Study Group

We developed a novel variation on classroom data collection by having students conduct a national research project. Students at 20 different colleges and universities measured “school spirit” at their institutions according to several operational criteria (school apparel wearing, car stickers, alumni donation rate, ratings by a major sports publication, and questionnaire measures). Instructors then combined this information into one large dataset, allowing students to analyze and compare trends measured at their school with those measured at other schools. We discuss the process of organizing a national study (recruitment of faculty participants, dissemination of instruments, compilation of data), aspects of the project that instructors thought were most educationally valuable, and substantive results of the study (how well the different measures of school spirit correlated).

Many instructors seek to stimulate an interest in empirical research by having students collect, generate, and analyze data in their classes. Instructors have implemented these exercises not only in research methodology and statistics courses, but also in courses such as close relationships and introductory psychology (Collins, 1997; Lutsky, 1986). Such activities appear to be based on the assumption that giving students greater “ownership” of their data will increase their intellectual investment in the principles of the course and, ultimately, their learning experience. According to Lutsky, there are three educational goals of empirical research projects: (a) increasing students’ familiarity with theories and phenomena in substantive content areas, (b) exposing them to tasks associated with empirical research, and (c) influencing “students’ attitudes toward the research claims of psychology” (p. 119).

We sought to apply this reasoning on a scale larger than the typical class project by linking a large number of individual class data-collection exercises into a national study. Specifically, classes at 20 colleges and universities throughout the United States carried out the task of measuring “school spirit” at their home institution. At least two aspects of the project carried the potential of enhancing students’ enthusiasm and intellectual commitment: (a) the fact that, as students at any one institution were gathering their data, they knew that other students around the country were working in parallel on the same project, possibly creating a sense of “team spirit,” and (b) students would likely have a strong interest in seeing how their school’s degree of spirit compared to that of other institutions.

We selected school spirit as the focal construct because it is an important reflection of who I am”) to which respondents indicated their degree of agreement or disagreement. The latter presented respondents with seven versions of a pair of circles—one circle always labeled “Self” and the other

Method

The project began with recruitment of faculty members to implement the study in their classrooms. The project originator (Alan Reifman) sent messages to two e-mail-based discussion lists (Society for Personality and Social Psychology and Evaluation, Measurement, and Statistics) in late July 2000 seeking participants for that fall. The request described the project as well as the incentive that participating instructors would receive a data file compiling results from participating schools to use in teaching statistical analysis.

Approximately 40 instructors replied to indicate at least a tentative interest in participating. Reifman then sent out instructions and measures for implementing the project (collaborators also made suggestions that led to modifications of the materials before distribution). On completion of data collection, each instructor e-mailed the data for his or her class to Reifman, who compiled the master database (there was no special data submission form or template; instructors submitted their data just by writing a regular e-mail message). Ultimately, 21 instructors at 20 institutions, representing approximately equal numbers of public and private schools, submitted data. (Where two instructors at the same university participated, we averaged their data, allowing schools to remain the unit of analysis.) Participating individuals’ names appear in the Appendix.

Classes obtained data in five areas at their schools: (a) percentage of students on campus who wore apparel depicting the home school’s name, insignia, sports teams, or other references (based on Cialdini et al., 1976); (b) percentage of cars in student parking lots that had school-related window decals, bumper stickers, or other items; (c) alumni giving rate, as obtained from the annual U.S. News & World Report college issue (America’s Best Colleges, 2000); (d) spirit associated with a school’s athletic program, as rated by The Sporting News (Gietschier, 1999); and (e) attitudinal measures of pride and commitment to the school, which students in the classes administered to other students around campus. The survey instrument for the attitudinal component adapted items from Luhtanen and Crocker’s (1992) Collective Self-Esteem Scale and Tropp and Wright’s (2001) visual measure of how close students perceived themselves to their schools. The former contained traditional self-report items (e.g., “[Name of school] is an important reflection of who I am”) to which respondents indicated their degree of agreement or disagreement. The latter presented respondents with seven versions of a pair of circles—one circle always labeled “Self” and the other
labeled “[Name of School] Students”—with the versions varying along a continuum from one endpoint where the circles completely overlapped (indicative of high closeness to the school) to the other endpoint where the circles were completely separate (indicative of low closeness).

Instructors of some classes divided students into task-specific groups, with field activities supervised by teaching assistants. As noted earlier, the unit of analysis for the final dataset was school. The rows of the data file represented schools and the columns contained the variables (measures of school spirit). Some of the variables came naturally in the form of a single figure per school (e.g., giving rate), whereas for others (e.g., the survey-based ones) each school’s representative needed to average multiple-respondents’ data to produce a single value.

Instructors needed to collect and send in their data as quickly as possible, so that the data would be available for teaching about statistical analysis in these courses. We believed statistical instruction would be more meaningful for students with data they had collected themselves than with, for example, data from a textbook or a “canned” software package. To give instructors maximum flexibility in when they could use multiple-school data for this purpose, Reifman used an “installment plan” to ensure the earliest distribution of the data. By early November 2000, instructors from 13 schools had reported their data to Reifman. He then sent an e-mail message to all instructors (even those who had not yet reported) containing the existing data in two formats: an SPSS for Windows file attachment and plain (ASCII) text within the body of the message. Each time, after the next few schools’ data arrived, Reifman e-mailed new cumulatively updated files, culminating with the final dataset.

Results

Two types of results are available from this exercise. The first consists of feedback, both quantitative and qualitative, from participating instructors on their perceptions of the exercise’s value. The second consists of analyses of the school spirit variables.

Feedback From Instructors

Among instructors who implemented the exercise, 17 responded to a postproject evaluation survey. The final sample size was 18, with the inclusion of a graduate teaching assistant for one of the responding professors. At the 17 institutions for which there were postproject data, the exercise implementation breakdown by type of class was 12 research methods classes (70.6%), one statistics class (5.9%), and four classes within mult course sequences encompassing statistics and methods (23.5%). The research methods courses were primarily within psychology departments, but also within those of sociology and human development and family studies. Class sizes ranged from approximately 15 at the low end to 85 to 100 at the high end.

The survey contained 17 evaluative dimensions. An anonymous reviewer suggested some of them, with the instructors contributing the rest. We phrased the dimensions in terms of ways the exercise may have contributed to students’ experiences, such as “Enhanced student engagement,” or “Aided development of students’ critical thinking skills” (see Table 1 for the full set of dimensions). Having the same people who implemented an exercise also do the evaluation clearly raises questions as to whether such an evaluation can be unbiased. In an attempt to reduce bias as much as possible, we refrained from asking instructors how much they agreed or disagreed with the dimensions. Instead, the task called for instructors to select five dimensions representing what instructors considered the most valuable aspects of the exercise. One could assume that the instructors, on the whole, probably were favorably inclined toward the exercise. However, this generalized positivity should not necessarily have biased anyone toward picking any particular dimensions over any other ones, thus leading to what we hope was a reasonably unbiased evaluation.

As shown in Table 1, some dimensions received strong consensus support: enhancing students’ understanding of course concepts such as operational definitions (with near unanimity), making the process of doing research accessible to students, and enhancing student engagement. Yet, there were instructors who found value in nearly all the nominated facets of the exercise; of the 17 dimensions, only four were not cited by at least two of the 18 respondents (i.e., did not achieve endorsement by at least 11% of respondents). We consider it a positive development that instructors found the exercise to be useful for many different purposes, although its

<table>
<thead>
<tr>
<th>Dimension</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enhanced students’ conceptual and practical understanding of course concepts (e.g., operational definition, convergent validity)</td>
<td>94</td>
</tr>
<tr>
<td>“Demystified” or made accessible to students the process of doing research</td>
<td>61</td>
</tr>
<tr>
<td>Enhanced student engagement</td>
<td>56</td>
</tr>
<tr>
<td>Enhanced student learning</td>
<td>39</td>
</tr>
<tr>
<td>Enhanced statistics teaching by enabling students to see where the numbers (data) come from</td>
<td>39</td>
</tr>
<tr>
<td>Could be integrated with other course material</td>
<td>33</td>
</tr>
<tr>
<td>Promoted collaborative teaching exercises among faculty members (in same or different schools)</td>
<td>28</td>
</tr>
<tr>
<td>Taught students practical research skills that may help them in the future (many jobs will require them to compile and synthesize information)</td>
<td>28</td>
</tr>
<tr>
<td>Was easy to implement/had a favorable cost-benefit tradeoff in terms of time and learning</td>
<td>28</td>
</tr>
<tr>
<td>Aided development of students’ critical thinking skills</td>
<td>28</td>
</tr>
<tr>
<td>Created class cohesion and a sense of collaboration among students</td>
<td>28</td>
</tr>
<tr>
<td>Cross-school aspect helped focus students’ attention on the importance of replication</td>
<td>17</td>
</tr>
<tr>
<td>Had value as a demonstration of active learning</td>
<td>11</td>
</tr>
<tr>
<td>Showed students how to learn in a community environment</td>
<td>6</td>
</tr>
<tr>
<td>Showed students how statistics tell a story</td>
<td>6</td>
</tr>
<tr>
<td>Allowed students to learn about a substantive topic of interest (school spirit)</td>
<td>0</td>
</tr>
<tr>
<td>Made lectures more “dynamic”</td>
<td>0</td>
</tr>
</tbody>
</table>

Note. N = 18. Each instructor could select five dimensions.
chief asset appeared to be its ability to promote the understanding of concepts. Given the exercise’s focus on measurement, it is not surprising that instructors saw value in the opportunity to teach about operational definitions, convergent validity, and related topics. The 39% citation rate for “Enhanced student learning” may strike some as low, given that learning is the major goal of teaching; however, this item had a very general wording and instructors may have focused in on the more specific ones.

Some of the instructors included unsolicited comments about the exercise when first submitting their classes’ data:

- “This has been excellent for my students.”
- “This was a neat project and well worth the effort.”
- “Thank you for allowing us to take part in this exercise. It made for much lively discussion in our laboratory sections!”
- “In addition to finding a neat result, the project has important pedagogical significance … I think the students became more interested in their project because they could realize the significance of the data. In addition, the distinction between units of analysis (individual for the class when each student generated data, and institution for the completed data set) allowed for interesting discussions about statistical methodologies.”

We also asked instructors about any negative aspects they observed about the project. One instructor, who obtained extensive feedback from her students, noted that negative reactions mostly reflected the dynamics of working with a group and not the project itself (e.g., perceived inequities in how much work group members did, difficulty coordinating times for groups to meet).

**School Spirit Data**

Correlations among the major indicators of school spirit appear in Table 2. All correlations were positive. Because schools were the unit of analysis (resulting in a small sample size), only the largest correlations were significant. Although we presumed all of the variables would tap into school spirit, we believe that the use of highly independent sources, sharing very little method variance, makes the correlations noteworthy. As one example, alumni giving rates correlated .78 with schools’ average responses to the visual closeness measure.

**Discussion**

The evaluative feedback from instructors suggested that this project increased student engagement and aided in substantive instruction about research methods and related topics. The “Internet Age” greatly facilitated implementation of this project. All communication between the organizer and participating instructors took place electronically through e-mail and attachments. By using list-serve discussion groups to recruit participants, hundreds of individuals were instantly reachable; lacking this feature, recruiting a sufficient number of participants might have been impossible. Although other excellent forums exist for collecting national data for teaching and research purposes, such as the Web-based Psych Experiments at the University of Mississippi (http://psychexps.olemiss.edu/) and the Implicit Association Test at Harvard University (http://implicit.harvard.edu/implicit/), the type of exercise we have described took students out into the field to collect different types of data (survey, observational, and archival), thus exposing them to the breadth of inquiry in psychological research. Based on our experiences, we believe that other classroom research projects that are national (or at least multicampus) in scope can achieve many of the same benefits as we observed. In the following paragraphs, we present suggestions for implementing multicampus projects in the future.

Any project will require a substantive topic. School spirit appeared to be a successful one for this exercise, but others could easily work. In the future, instructors and their classes could apply the same basic methods to the operational definition and measurement of other constructs. One of the participating instructors extended this measurement approach in a later class by having students measure patriotism after the September 11, 2001 attacks. Other suggestions include conducting attitude surveys or trying to replicate a classic psychology experiment across campuses. In addition to the teamwork aspect of including multiple campuses, doing so can also promote classroom discussion of the importance of replication and the idea of moderator variables (school-level characteristics that may explain why different classes obtained different findings).

Other important considerations revolve around the degree to which instructors must deliver the exercise to students in a “top-down” fashion (i.e., with the ideas and procedures already determined), and how much instructors will be able to incorporate student input into study design. As noted earlier, giving students greater ownership of the project may increase their engagement in the activity. However, doing integrative and comparative data analyses across schools requires substantial uniformity in study design across classes. Achieving this uniformity while including student input into study design obviously presents a major challenge. Time availability and the number and location of collaborating instructors would appear to be important factors. Unless one uses a full-year course sequence, instructors and students must accomplish many tasks within the timeframe of a single academic term; these include introducing the exercise to students, dividing the tasks, collecting data in the field, entering data into the computer, conducting analyses, and in some cases having students write up

<table>
<thead>
<tr>
<th>Table 2. Correlations Among Measures of School Spirit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decals and stickers</td>
</tr>
<tr>
<td>Collective self-esteem (total)</td>
</tr>
<tr>
<td>Visual closeness measure</td>
</tr>
<tr>
<td>Giving</td>
</tr>
</tbody>
</table>

*Note. Sample sizes ranged from 16 to 19 schools. **p < .001, (two-tailed).
APA-style reports. If classes are to accomplish all these tasks in a single term while following a uniform research design, it would appear necessary in many cases for the instructors to have a “prepackaged” study design ready to present to students at the beginning of the course.

Such a conclusion might be less applicable, however, the smaller the scope of the project. If, for example, instructors at three institutions within the same city wanted to conduct a multicampus project, it would be possible to have a joint introductory meeting of all three (or however many) classes. The presence of all instructors and students in one location would aid in the expeditious consideration of all suggestions for study design, including those from students. A small-scale study would thus allow all the desired features to be present: student input, a uniform research design, and quick implementation.

Even if centralized, top-down delivery of the study design to students appears necessary, there may still be the opportunity for “local variation” that would encourage greater student input. Classes would adhere closely to the basic research design, but instructors and students could branch off in different directions once they had collected the basic data. Some of the present instructors did just that by, for example, incorporating other variables for which data about the colleges were publicly available, generating additional hypotheses, and conducting additional types of statistical analyses. Also, although this report has focused on analyses with school as the unit of analysis, each instructor had the opportunity to have his or her class conduct individual-level analyses using the questionnaire data. The idea of local variation thus allows each instructor to tailor the conceptual framework and data-analysis strategy of the exercise to his or her interests and the needs of the course, while preserving the uniformity of study design needed to build the larger multischool dataset.

Having sufficient time once classes start to include student input and quickly initiate the project is not the only temporal issue. The nature of the project will almost certainly be affected by how far in advance of the beginning of classes the instructors come together (electronically, if not physically) and begin planning the exercise. In this case, the organizer sent out the call for collaboration approximately 1 month before the start of fall classes at most institutions. Because of the closeness to the start of classes, the recruitment message already contained the basic study framework, and collaborators could only suggest modifications. Clearly, the earlier instructors come together, the more they can discuss study design and come up with more elaborate features (e.g., experimental manipulations) than those of the present project, if desired.

One negative aspect, as noted earlier, involved students’ difficulties in working with group-mates. As a partial aid to readers who will be conducting group-based activities in the future, Hoffman and Rogelberg (2001) published a study showing college students’ preferences for how group projects should be graded (i.e., their judgments of the optimal degree to which individual input and overall quality of the group project should be weighed in grading).

In conclusion, a little creativity on the part of instructors, combined with the tremendous capabilities of electronic communication, can facilitate development of multicampus projects. Earlier writers have argued that research-based activities enhance students’ learning experiences, and this type of exercise may increase the educational benefits further still.

References


Appendix

Participating Institutions and Investigators in the School Spirit Study Group

Alabama-Huntsville, Sandra Carpenter; Baylor (TX), Tamara Rowatt; Buffalo State (NY), Lisa Brooks (with Victoria Magid, Robert Stage, and Paulette Wydro); Georgia, Steve Cramer; Gustavus Adolphus (MN), Marie Walker; Hanover (IN), Connie Wolfe; Holy Cross (MA), Royce Singleton; Maryland-College Park, Harold Sigall (with Angela Eichelberger, Julie Jordon, and Samantha Leaf); Monmouth (IL), Jon Grahe; Oklahoma, Ryan P. Brown; Pennsylvania State, Janet Swim (with Nicholas B. Pearson); Rhodes (TN), Chris Wetzel; South Florida-St. Petersburg, Mark Pezzo; Texas-Austin, Sam Gosleng; Texas-El Paso, Kim MacLin; Texas Tech, Alan Reifman* (with Brandon Awbrey and Collyn Wright); Trinity (TX), Page Jerzak; U.S. Air Force Academy (CO), Steven M. Samuels; Virginia Tech, Greg Lemmond; Wake Forest (NC), Mark Leary and Catherine Seta.

*Organizer (Names in parentheses refer to teaching assistants.)

Notes

1. Steven L. West made major contributions to adapting the Collective Self-Esteem Scale to college affiliations in his dissertation work.

2. The School Spirit Study Group consisted of investigators from 20 colleges and universities throughout the United States. The organizer of the group is Alan Reifman, from whom copies of the instructions and measures used in the exercise are available.

3. Send correspondence to Alan Reifman, Department of Human Development and Family Studies, College of Human Sciences, Texas Tech University, Lubbock, TX 79409–1162; e-mail: alan.reifman@ttu.edu.
Demonstrating the Gambler's Fallacy in an Introductory Statistics Class

Todd C. Riniolo and Louis A. Schmidt

*Teaching of Psychology* 1999 26: 198
DOI: 10.1207/S15328023TOP260308

The online version of this article can be found at:
http://top.sagepub.com/content/26/3/198
In this article, we describe a classroom demonstration that uses the Gambler’s Fallacy to illustrate misconceptions about random processes and how they affect statistical interpretation. The demonstration used a database collected from simulated gambling by students picking professional football games with the point spread (i.e., a real-life random process). The results of student picks illustrated that random processes are not self-correcting and reinforced the relation between sample size and variability. Formal and informal feedback from students indicated that the demonstration was well received and recommended for future classes.

People often have inaccurate perceptions about random processes (Bar-Hillel & Wagenaar, 1993). For example, basketball fans and players tend to believe the chances of an individual player scoring a basket are greater if the player was successful in making previous shots (i.e., streak shooting). However, streak shooting is a misconception, as previous shots provide no reliable information about the success of subsequent shots (Gilovich, Vallone, & Tversky, 1985).

Such misconceptions about random processes contribute to students being poor intuitive statisticians (Tversky & Kahneman, 1971). Specifically, many students possess an overconfidence in the stability of results obtained from a single small sample (Tversky & Kahneman, 1974). Tversky and Kahneman (1974) speculated that students generalize the precision of large samples to small samples because of an erroneous belief that random processes are self-correcting (i.e., errors cancel each other out). For example, people often infer a
greater probability of a coin landing tails when preceded by several consecutive heads (Bar-Hillel & Wagenaar, 1993). This misconception implies an expectation that even short sequences (or small samples) will self-correct chance deviations. This generalization from large to small samples occurs despite learning the standard error formula that shows the relation between sample size and statistical precision. Thus, students may not routinely retain a basic statistical principle (i.e., the relation between sample size and sampling variability) because of a misinterpretation about the influences of random processes on statistical outcomes (Tversky & Kahneman, 1974).

Unfortunately, there are limited teaching demonstrations available that address the role of random processes on statistical interpretation. Several variations of regression toward the mean demonstrations are available to illustrate the influences of chance contributions to extreme scores (e.g., Karyowski, 1985; Levin, 1982). However, these demonstrations imply a "corrective mechanism" toward the mean, without an explanation for students about how random processes truly function (i.e., no self-correction). Furthermore, these demonstrations did not address how random processes can alter group results or the relation between sample size and variability. Thus, additional demonstrations would add to the existing literature.

In this article, we describe a classroom demonstration performed in an introductory psychology statistics class. The demonstration assumed no previous background in statistics and presented the material in an interesting format to enhance learning (Dillbeck, 1983). Simulated gambling using weekly results from class selection of professional football game outcomes (with the point spread) created the database for the demonstration. We chose this format because Las Vegas point spreads for football are designed to elicit equivalent amounts of money wagered on both teams as profits are made by not paying off at a one to one rate (i.e., winners receive slightly less money than losers pay). Research demonstrates that individuals have a 50/50 chance of correctly picking games with the point spread (Canes, 1974). Thus, simulated wagering on football games provides one method for students to actively create and observe a real-life random process. Specifically, this article describes the Gambler's Fallacy demonstration designed to illustrate that (a) random processes do not self-correct and (b) statistical precision is sample size dependent.

Method

Participants

Twenty-five women (13 self-identified football fans) and 20 men (18 self-identified football fans) completed a brief survey during the first week of the semester. Of the 45 students, all but 2 expressed an interest in learning about the probabilities associated with gambling.

Procedures

Class introduction. Early in the semester, we announced to our class that they would pick 12 professional football games (with the point spread) weekly over a 10-week period. We told the students that picking games would be an ongoing project throughout the semester that would result in a database used to teach statistical lessons. To put the class at ease, we explained that the demonstrations would be user friendly, would not require individuals to be football fans, and would be related back to social sciences research. The picks were made on the last class period prior to the weekend and took approximately 5 min to complete.

The Gambler's Fallacy demonstration. We asked students to answer the question in the following scenario approximately halfway through the semester:

Throughout the semester, you discover that you have a 50/50 chance of correctly picking football games with the point spread. During your vacation in Las Vegas over Thanksgiving break, you decide to wager (bet) on football. Unfortunately, you lost the first three games on which you wagered. Approximately what percentage chance do you have to correctly pick the next game?

After students answered the question, a discussion began based on the classic article by Tversky and Kahneman (1971), which showed how common misconceptions about the laws of chance relate to research interpretation in the social sciences. Of the 43 students who participated, 28 responded with the correct answer (i.e., 50%). The discussion centered around whether a random process is self-correcting (i.e., do errors from a random process "cancel each other out?"). We ended the discussion by explaining what the implications would be if random processes were self-correcting. For the previous scenario, it implies (a) that the gambler (after correctly picking or missing three games in a row) can affect the outcome of pro football games often played hundreds or thousands of miles away by virtue of his or her next pick, (b) that a fair roulette table can remember if previous spins were black or red, or (c) as Tversky and Kahneman pointed out, that coins have a memory to balance out heads and tails.

To demonstrate the Gambler's Fallacy that there is no corrective bias associated with random processes, we used the existing database of football picks. We described to the class that correctly picking or missing three games in a row should not influence the chances of the next pick being right or wrong. We predicted that the results should be close (allowing for measurement error) to the class average of picking games if random processes are really not self-correcting.

We tallied and shared the results with the class the following week. At this point in the semester (5 weeks worth of picks), the class was picking games correctly 53% of the time. Results showed that (n = 256 occurrences) after correctly picking three games in a row, the chances of picking the fourth game correctly were 52%. Likewise, after missing three games in a row (n = 200), the class average for picking the fourth game correctly was 54%. These results (within 1% point of the class average of 53%) reinforced the previous discussion that random processes are not self-correcting. Thus, outcomes of previous games provide the gambler with no meaningful information for future wagers.

During the same class period, a discussion followed the Gambler's Fallacy demonstration focusing on the effects of
varying sample sizes in social sciences research. The discussion emphasized that because random processes (i.e., random sampling in social sciences research) are not self-correcting, small sample sizes are at a qualitative disadvantage compared with larger sample sizes. Larger samples can dilute the effects of deviations or extreme scores as sampling proceeds, resulting in more reliable and precise findings. The discussion led to questions by students about the appropriate sample size in social sciences research. (Note: Do not perform the previous demonstration until acquiring a large database of scores to dilute deviations.)

Also, at the end of the 10-week period (5 weeks since the Gambler’s Fallacy demonstration), we presented the final results to reinforce the previous discussion by providing students with a concrete example of the increased stability that accompanies increased sample size. Our a priori prediction (i.e., picking individual games with the point spread is 50/50) closely matched our a posteriori results as students chose individual games correctly 50.8% of the time (n = 382 total games selected over 10 weeks). Also, we presented our class with the average percentage of correct picks for each individual week. Out of the 10 total weeks, the class average exceeded 50% for 4 out of 10 weeks and ranged from a high of 58.2% (n = 41 students) to a low of 43.1% (n = 39 students). Presenting the final results allowed an additional opportunity to reemphasize the differences in variability between large (i.e., the total 10-week period) and smaller (i.e., each individual week) samples.

Student Evaluation

We examined students’ perceptions (n = 35) of the classroom demonstration using a 1 (strongly disagree) to 5 (strongly agree) metric. A median score of 4 for all evaluation questions indicated that most found the demonstration to be enjoyable (SD = .73), contributed to their overall understanding of statistics (SD = .87), preferred the demonstration format to straight lecture (SD = .75), and recommended that this demonstration continue for future classes (SD = .78). These results were consistent across men and women and football fans as well as non-football fans.

Discussion

Tversky and Kahneman (1971) maintained that a warning about the misleading nature of random processes and how random processes may affect statistical interpretation should accompany a student’s introduction into statistics. Our demonstration provides instructors with a vehicle to accomplish that goal in an interesting format. Specifically, the Gambler’s Fallacy demonstration is useful to illustrate to students that real-life random processes are not self-correcting and to reinforce the relation between sample size and variability.

In addition, many students commented that after the discussion of what the implications would be if random processes were self-correcting (e.g., a coin has a memory), the Gambler’s Fallacy made sense. However, research indicates misconceptions about random processes often survive despite considerable contradictory evidence (Bar-Hillel & Wagenaar, 1993). The effectiveness of our demonstration over time is unknown, but it may have alerted students to (a) recognize and take precautions against their own misconceptions and (b) not generalize the statistical precision that accompanies large samples to small samples.

Finally, instructors may be concerned with whether simulated gambling is appropriate for the classroom. We believe this format was a way to keep interest high throughout the semester, and we were careful to stress the realities of gambling that our demonstrations helped to emphasize including (a) no advantages exist in the short run based on past events (e.g., the Gambler’s Fallacy) and (b) the gambler loses over the long run because Las Vegas casinos do not pay off at a one to one rate (e.g., the relation between sample size and variability). Finally, for instructors who wish to perform the Gambler’s Fallacy demonstration during the Spring semester, simply utilize alternative (e.g., basketball, hockey) point spreads.

References


Notes

1. Special thanks to Marv Levy for his helpful comments.

2. Send correspondence to Todd C. Riniolo, Department of Psychology, Adams State College, Alamosa, CO 81102; e-mail: triniol@adams.edu.
Demonstration of Factors Affecting the $F$ Ratio

Mark J. Sciutto  
Westminster College

This demonstration provides an interactive, nonquantitative approach to understanding factors contributing to the $F$ ratio in a one-way ANOVA. The demonstration is relatively easy to use and requires minimal setup and preparation. Students made decisions about the relative weights of 2 boxes, 1 representing between-groups variance and the other representing within-groups variance. The contents of the boxes were systematically manipulated to demonstrate the concepts of power, effect size, and errors in hypothesis testing. Students in 2 introductory-level statistics courses perceived the demonstration as very useful and recommended use of the demonstration in the future.

Johnson (1989) presented an intuitive demonstration of the $F$ ratio that is relatively simple and involves only basic numerical manipulation. In his demonstration, Johnson introduced the concepts of between- and within-groups variance by systematically manipulating raw data and observing the changes in the ANOVA summary table. According to Johnson, the use of concrete in-class demonstrations is potentially valuable to students, particularly those who have difficulty with the highly abstract nature of many inferential procedures.

The demonstration described here is an extension of Johnson’s (1989) demonstration. In addition to comprehending the distinction between the two sources of variance, it is also desirable for students to gain an understanding of the various factors contributing to these sources (i.e., treatment effect, individual differences, and measurement error). I have used the following demonstration to help students systematically experience the influence of various factors on the $F$ ratio with minimal use of numbers and formulas. The goal of the demonstration is to promote a greater understanding of how effect size, individual differences, and measurement error combine to influence the ability to detect a false null hypothesis (i.e., power) in a one-way ANOVA.

Method

Materials

I used two black lightweight cardboard boxes (approx. 8 cm × 18 cm × 13 cm) for this demonstration. One box was labeled “between” and the other was labeled “within.” Prior to the demonstration, I selected assorted small objects of variable weights (e.g., stopwatches, action figures, batteries) to place in the boxes.

Procedure

After an introductory discussion of the one-way ANOVA, I presented Johnson’s (1989) demonstration. Following a discussion of the difference between within-groups and between-groups variance, I introduced the various factors that could affect each of these sources of variance (i.e., What would cause one group to differ from another? What would cause people within the same group to differ from each other?). I then passed each of the small objects around the class and declared that certain objects (e.g., small action figures) represented individual differences, other objects represented measurement error (e.g., stopwatches), and the remaining objects (e.g., different size batteries) represented various treatment effects. I chose these specific objects because they seemed to have an intuitive connection to what they represented.

After the class inspected the objects, I selected a student volunteer to stand in front of the class and close his or her eyes. After placing one box in each of the student’s hands, I then asked whether he or she could tell if there was a difference in the weights of the 2 boxes. The student also estimated his or her level of certainty (0 to 100%) regarding the existence (or nonexistence) of a difference. We repeated this procedure several times, while I systematically manipulated the content of the boxes as follows:

Case 1. Each box contained only measurement error (two stopwatches) and individual differences (four action figures).

Case 2. I added a small treatment effect (e.g., one “AAA” battery) to the “between” box.

Case 3. I added a medium treatment effect (e.g., one “C” battery) to the “between” box.

Case 4. I added a very large treatment effect (e.g., two “D” batteries) to the “between” box.

Case 5. I removed nearly all or all of the individual differences and measurement error from both boxes. I then repeated each of the previous steps.

In general, students’ ratings of certainty correlated positively with the addition of larger treatment effects and with the removal of random error. For example, student ratings of certainty ranged from 10% to 30% for Cases 1 and 2, but were always near 100% for Case 4. After the demonstration, I asked the class to describe how adding larger treatment effects influenced the person’s ability to accurately detect a difference in the weights of the two boxes. We also discussed how reducing random error (i.e., individual difference and measurement error) led to more accurate detection of differences in the weights of the two boxes. To conclude the demonstration, I explicitly reviewed the connection between the demonstration results and the concepts of effect size (i.e., adding batteries), power (i.e., ability to accurately detect differences), and errors in hypothesis testing (e.g., incorrectly identifying a difference).

Evaluation

I used this demonstration in two introductory Research Methods and Statistics courses. There were 16 students in the first section (Fall 1996) and 25 students in the second section (Fall 1997). After the demonstration, students anonymously evaluated the usefulness of the demonstration. Specifically, students rated the usefulness of the demonstration...
in learning the concepts from 1 (not at all useful), to 7 (extremely useful) and whether they would recommend using the demonstration in the future (yes or no). The mean rating of usefulness for the sample (N = 41) was 5.79 (SD = .93), with 66% of the students giving a rating of 6 or 7. When asked whether I should use the demonstration again, 40 of the 41 students responded “yes.” In sum, I have found this demonstration to be a useful tool for illustrating the complex interrelation of effect size, random error, power, and errors in hypothesis testing.

Reference


Note

Send correspondence to Mark J. Sciutto, Department of Psychology, Westminster College, New Wilmington, PA 16172; e-mail: sciutt mj@westminster.edu.

Floating Data and the Problem With Illustrating Multiple Regression

Daniel A. Sachau
Minnesota State University, Mankato

In this article, I describe a technique for introducing students to multiple regression using an easily constructed, large-scale, three-dimensional model. The model helps instructors explain multivariate normal distributions, centroids, Mahalanobis distance, multivariate outliers, and regression.

It is relatively easy to illustrate simple regression at a chalkboard. Draw a horizontal X-axis predictor, a vertical Y-axis criterion, a few data points, a regression line, and there it is. Illustrating regression becomes considerably more difficult, however, when you add an additional predictor. Instructors who discuss multiple regression often ask students to imagine another axis emerging out from the chalkboard. The instructors then introduce the idea of data points with three, rather than two, coordinates and ask students to imagine the data points “floating” out in the classroom. Once data points leave the chalkboard, students often begin to struggle with demonstrations of regression.

In an effort to introduce the basic concepts of multiple regression, I create a large-scale, three-dimensional regression model using the classroom walls and floor. That is, instead of using imaginary axes, I create concrete axes by making the length of the right wall correspond to an X-axis predictor, the length of the adjoining wall correspond to a Z-axis predictor, and the height of the walls correspond to the Y-axis criterion. Then I bring in data points.

I use fishing bobbers to represent data points. After removing the furniture from half the classroom, I use fishing line and thumbtacks to suspend 50 red and white bobbers from the ceiling so that the pattern of bobbers represents three positively correlated and normally distributed variables. In other words, I arrange the bobbers in the shape of a large football. The football is tilted so that one of the points of the football is near the floor in the corner of the room (the junction of the X-, Y-, and Z-axes). The other end of the football is near the ceiling at the center of the rectangular room.

Next, I print large letters (X, Y, Z) and three sets of numbers (1 to 7) and tape them to the walls. If the classroom is constructed from concrete blocks, the joints of the blocks provide convenient points for locating axes coordinates. It takes about 30 min to set up the demonstration.

The bobber model is similar to a model described by Kirby (1976). Kirby's model involved pegboard, rods, and spheres. Although less portable than Kirby’s model, the bobber model allows students to walk inside the distribution.

I use an employee selection example to start the discussion of multiple regression. I ask students to choose a job, a performance criterion, and two predictors. For simplicity, I ask students to select, or make up, variables with a 7-point range. For example, students might try to predict supervisory ratings of a clothing store manager’s performance using years of retail experience and assessment center performance as the predictors. Students usually come up with much more interesting or exotic examples.

Teaching Points

Centroid

The centroid is the point defined by the means of the three variables. It floats in the center of the football. For illustration, almost any bobber that is different from the data point bobbers will work nicely as a centroid.

Multivariate Normality

I ask students why three normally distributed and positively correlated variables take the shape of a football. Students have to think about how the normally distributed variables, each of which is thicker in the middle than it is on the ends, combine to form the football shape.

Mahalanobis Distance

Mahalanobis distance is the distance between a data point and the centroid (a bobber and the unique bobber). Like the centroid, Mahalanobis distance is tricky to explain but easy to point to. Mahalanobis distance is useful for defining an outlier.

Multivariate Outlier

Outliers can be a problem because outliers have a disproportionately high degree of leverage on the position of the regression line. Outliers, thus, may need to be eliminated. A
A Sweet Tasting Demonstration of Random Occurrences
Andrew N. Christopher and Pam Marek
Teaching of Psychology 2002 29: 122
DOI: 10.1207/S15328023TOP2902_09

The online version of this article can be found at:
http://top.sagepub.com/content/29/2/122
A Sweet Tasting Demonstration of Random Occurrences

Andrew N. Christopher
Anderson College

Pam Marek
St. Michael's College

We describe an active learning demonstration using LifeSavers® candy to illustrate the phenomenon of random occurrences. We provide each student with a roll of Five Flavor LifeSavers. In a dyad, students take turns trying to guess the flavor of candy in their mouths when deprived of the senses of sight and smell. Because there are 5 flavors of candy, students have a 20% chance of randomly guessing the correct flavor. In a sample of 49 undergraduates, student performance reflects this chance guessing rate. Students report that this demonstration was informative and enjoyable. Exam results indicate that students mastered the concept of random occurrences. We discuss other uses of this demonstration.

Understanding of chance or random occurrences is one facet of critical thinking that underlies interpretation of scientific experimentation. On a broader scale, misperception of randomness is one of several difficulties in statistical or probabilistic thinking that may explain how people know what is not so (Gilovich, 1991). Specific descriptions of the misperceptions of randomness and related biases in statistical thinking (e.g., illusory correlation or the gambler’s fallacy) are highlighted in only about half of the Introductory Psychology texts published between 1995 and 1997 (Griggs, Jackson, Marek, & Christopher, 1998). How might instructors introduce this important topic in a manner that will engage students who may otherwise consider statistics a somewhat distasteful subject area?

Concrete, hands-on experiences may help increase understanding of more abstract statistical concepts (Brown & Kane, 1988). To this end, Dyck and Gee (1998) discussed how teach-
ers can have students use M&M’s® candy to create an actual sampling distribution of the mean for a designated candy color. Similarly, Hull and Hull (2000) flavored their presentation of subjective, empirical, and theoretical probabilities by distributing a bag of plain M&M’s to each student and then actively involving students in estimations and statistical comparisons of the probability of finding each of six M&M’s colors. In this article, we describe a demonstration in which students learn about probability by attempting to guess the flavor of a LifeSavers® candy. Because class discussion subsequent to the activity can potentially extend in multiple directions, we consider how to incorporate this sweet tasting demonstration into a range of psychology courses.

Conducting the Demonstration

Materials and Preparation

LifeSavers, and possibly small sandwich bags, are the only materials required for this demonstration. The choice of individual rolls or larger bags of LifeSavers depends on class size and instructor preference. One alternative is to acquire one roll of Five Flavor LifeSavers for each student. Each roll contains the following flavors: cherry, lemon, lime, orange, and pineapple. Another alternative is to buy large Five Flavor packages of individually wrapped LifeSavers (approximately one package for every five students) and a sufficient number of small sandwich bags to enable pairs of students to share a bag. To prepare for the activity, the teacher should place two LifeSavers of each flavor in a sandwich bag. A third alternative is to “stock up” on economy sized bags of the candy sold around the Halloween and Easter holidays.

Procedure

Announce that the class will be playing a guessing game with LifeSavers’ flavors. Write the names of the five flavors on the board (or use an overhead transparency prepared in advance). Ask students to form dyads (if there is an odd number of students, the teacher can participate also). Indicate that the object of the game is to guess the flavor of a LifeSaver in one’s mouth without using the senses of sight or smell. After reminding students again about the five available flavors, distribute a roll of LifeSavers to each student (or a sandwich bag of LifeSavers to each dyad). Emphasize that students should not open the roll (or the bag) until the instructor explains the following guidelines.

Students in each pair take turns eating a LifeSaver and guessing its flavor. Those guessing first close their eyes and tightly hold their noses. The partners of these students then open the roll of candy (or remove one piece from the sandwich bag) and place one unwrapped LifeSaver in their partners’ free hand. The students designated to guess first then eat the LifeSaver, attempt to determine its flavor, and report the flavor to their partners. Students then reverse roles and repeat the process.

After all students have an opportunity to guess the flavor, ask those who correctly identified the flavor to raise their hands. Count the number of correct identifications and convert this number into a percentage. (We have found that many students want a second chance to do this demonstration. Depending on time constraints, the teacher may allow students to repeat the process. If the process is repeated, we suggest separate hand counts, one after each time both students have completed the procedure.) In a class of 35 students, this demonstration required about 10 min of class time to explain and execute.

Alternative Procedures

In larger classes, if the mechanics of the taste test approach appear cumbersome, instructors may opt for simpler, albeit less engaging, alternatives. In one alternative procedure, materials are limited to a single roll of LifeSavers. After indicating the five flavor designations, the instructor displays an unopened roll of LifeSavers, then asks all students to write down the name of the flavor they expect to appear first when the roll is opened. The instructor then recruits a student volunteer to open the roll, announce the first flavor to the class, and assist in tallying the count of students who have guessed correctly. Because this revised procedure does not involve eating candy, a particularly appealing aspect of the activity, from a student perspective, is lost. Thus, we suggest a compromise—an alternative procedure in which students guess the flavor they expect to appear, but also have an opportunity to eat the LifeSavers. In this scenario, the instructor distributes one unopened roll of LifeSavers to a group of five students, asks each student to guess the flavor that will first appear when the roll is opened, and appoints one individual in each group to tally the number of correct answers. The instructor then cumulates the group responses and announces the results to the class. During the subsequent discussion, students in each group may share the LifeSavers used in the demonstration. Although this alternative procedure may be attractive for larger classes, we have not actually used it ourselves, and thus cannot comment on its effectiveness.

Demonstration Results and Classroom Discussion

Results

Because this exercise restricts the use of both vision and smell (needed in combination with taste to detect flavor), it should not be possible for students to detect the flavor of candy in their mouths. Yet, about 20% of the students were still able to guess the correct flavor. Of the 49 students participating in this demonstration (35 in Introductory Psychology and 14 in Research Methods), 10 (20.41%) correctly guessed the flavor of LifeSaver in their mouths.

After counting the number of students who correctly identified the flavor of LifeSaver (or who correctly guessed the first flavor in a roll), begin a discussion of the demonstration results. Ask why certain students correctly identified the flavor. We have found that very few students specifically noted that given five flavors from which to choose, approximately
20% would guess correctly by chance. Instead, students who correctly identified the flavor of candy tended to emphasize their “superior taste buds.” Other students mentioned ways that these students may have cheated, leading to a discussion of control. In fact, the idea of having partners actually record responses in writing stemmed from student suggestions. Additionally, to ensure total restriction of sight and smell, students suggested the use of blindfolds and nose clips.

Applications and Potential Topics for Discussion

When students understand how probabilities affected the results of this demonstration, the subsequent direction of class discussion depends on course context. For example, in Research Methods courses, the instructor might extend discussion to hypothesis testing in research, encompassing such issues as what a $p$ value actually measures. Depending on class level in the curriculum, teachers can expand coverage to encompass how the concept of probability relates to random sampling. If two students shared a single sandwich bag of LifeSavers, probes concerning differences between sampling with and without replacement are appropriate. Why, for example, does the theoretical probability for a correct guess differ for the first and second taste testers in a dyad?1

In courses other than Research Methods, different avenues of discussion are suitable. In Introductory Psychology, teachers may link the demonstration to sensation and perception (e.g., sensory interaction), thinking (e.g., misperceptions of randomness), or the scientific attitude (e.g., the role of chance factors in assessing controversial claims related to extrasensory perception). In Cognitive Psychology, the demonstration can be used to complement material on decision making, such as how the representativeness heuristic contributes to misperceptions of randomness.

Evaluation of the Demonstration

We asked students ($N = 49$) to assess this demonstration by responding to five items taken from Hull and Hull (2000) on a 7-point scale ranging from 1 (strongly disagree) to 7 (strongly agree). Students reported that this demonstration was useful ($M = 6.12, SD = 1.34$), helpful in understanding random occurrences ($M = 6.00, SD = 0.91$), easier to learn from than from descriptions in the book ($M = 6.34, SD = 0.67$), a good supplement to the lecture ($M = 5.54, SD = 1.95$), and suitable for use in future classes ($M = 6.55, SD = 0.82$). None of these ratings differed significantly between the Introductory Psychology and Research Methods students.

We also evaluated the exercise by examining responses to related exam questions that we gave students approximately 1 to 2 weeks after the demonstration. We asked students in both the Introductory Psychology and Research Methods courses the following two multiple-choice questions, adapted from Brink (1998):

1. In a series of six coin flips, which of the following sequences of heads (H) and tails (T) is MOST LIKELY to occur?
   a) HTHTHT
   b) HHHHTT
   c) TTTTTT
   d) All of the above are equally likely to occur.*

2. The King James version of the Bible was completed when William Shakespeare was 46 years old. In Psalm 46 of this translation, the forty-sixth word is “shake,” and the forty-sixth word from the end is “spear.” Before concluding that the biblical translators were trying to be humorous with these word placements, you would be best advised to realize the dangers of:
   a) explaining events in hindsight.
   b) generalizing from small samples.
   c) perceiving order in coincidental events.*
   d) assuming people share your opinions.

*Indicates correct answer.

Of the 35 Introductory Psychology students, 28 correctly answered the first question, and 32 correctly answered the second question. Of the 14 Research Methods students, 12 correctly answered the first question, and all correctly answered the second question. Given the applied nature of these questions, these correct response rates suggest that this demonstration encourages students to consider the role of randomness in novel situations.

We acknowledge the absence of a control group that did not experience the demonstration. Furthermore, the fact that we assigned reading on this topic and lectured on it after doing the demonstration may have further enhanced performance on these exam questions. However, the level of correct responses hints that the demonstration indeed possesses pedagogical value.

In addition to the two multiple-choice questions, we asked the Research Methods students the following short-answer question: “In class, we did an exercise that required you to identify a flavor of candy when your sight and smell (essential for detecting flavor) were restricted. Explain how approximately 20% of the class managed to identify the flavor.” All 14 students in the class articulated how chance influenced the results of this exercise.

Conclusions

Enthusiastic student feedback and accuracy of students’ responses to exam questions indicate that the LifeSavers demonstration is an effective means to explicate how random

---

1Gravetter and Wallnau (1999) indicated that in a random sample “(a) each individual in the population must have an equal chance of being selected, and (b) if more than one individual is to be selected for the sample, there must be constant probability for each and every selection” (p. 135). In the sandwich bag version of this demonstration, the probability of selecting any particular flavor is 2 out of 10 for the first “guesser.” However, the probability for the second guesser is altered (conditioned) because there are only 9 LifeSavers remaining: 2 of 4 flavors and 1 of the 5th.
occurrences can be perceived as orderly. The demonstration tastefully offers an enjoyable presentation of the concept of probability and discussion is fruitfully channeled into other areas related to the scientific attitude, sensation and perception, and biases in decision making. Thus, the activity appears to be a useful addition to a variety of psychology courses at different levels of the curriculum.

References


Notes

1. Pam Marek is now at Anderson College.
2. We thank Cynthia S. Koenig and Jason R. Jones for their helpful comments on earlier drafts of this article.
3. Send correspondence to Andrew N. Christopher, who is now at the Department of Psychology, Albion College, Albion, MI 49224; e-mail: achristopher@albion.edu.
Enhancing Students’ Ability to Use Statistical Reasoning with Everyday Problems
Timothy J. Lawson, Michael Schwiers, Maureen Doellman, Greg Grady and Robert Kelnhofer
Teaching of Psychology 2003 30: 107
DOI: 10.1207/S15328023TOP3002_04

The online version of this article can be found at:
http://top.sagepub.com/content/30/2/107

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Society for the Teaching of Psychology

Additional services and information for Teaching of Psychology can be found at:

Email Alerts: http://top.sagepub.com/cgi/alerts
Subscriptions: http://top.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Apr 1, 2003

What is This?
Enhancing Students’ Ability to Use Statistical Reasoning With Everyday Problems

Timothy J. Lawson
Michael Schwiers
Maureen Doellman
Greg Grady
Robert Kelnhofer
College of Mount St. Joseph

We discuss a technique for teaching students everyday applications of statistical concepts. We used this technique with students (n = 50) enrolled in several sections of an introductory statistics course; students (n = 45) in other sections served as a comparison group. A class of introductory psychology students (n = 24) served as a second comparison group. We assessed students’ statistical reasoning ability at the beginning of the semester as well as later in the semester. All 3 groups showed improvements in statistical reasoning, but the greatest improvement occurred in the group that read the everyday application material.

For years, psychologists have discussed the importance of teaching students everyday statistical reasoning skills and the fact that typical statistics courses often fall short of this goal (e.g., Nisbett, Krantz, Jepson, & Kunda, 1983). Fong, Krantz, and Nisbett (1986) wrote:

If introductory statistics courses were to incorporate examples of how statistical principles such as the law of large numbers can be applied to judgments in everyday life, we have no doubt that such courses would have a more far-reaching effect on the extent to which people think statistically about the world. (p. 282)

This quotation evokes the question of which statistical concepts instructors should emphasize when teaching applications to everyday problems. Concepts that are relevant to everyday decisions but difficult for people to grasp or apply are a good starting point. We discuss suggestions for teaching about five such concepts: probability, the law of large numbers (LLN), estimation and sample bias, correlation, and regression toward the mean.

After teaching some basic rules of probability, professors might discuss the role of heuristics in producing faulty subjective probabilities. For example, the availability heuristic might cause people to overestimate the frequency of violent causes of death (e.g., Combs & Slovic, 1979), the anchor-adjustment heuristic may cause an overestimate of the probability of conjunctive events (Paulos, 1988; Tversky & Kahneman, 1974), and the representativeness heuristic might lead to the conjunction fallacy (Tversky & Kahneman, 1983). Underestimation of the probability of chance events is another important topic (e.g., Paulos, 1988; Stanovich, 2001). With respect to the LLN, professors could teach students about violations of this rule related to (a) people’s overreliance on single-case evidence (e.g., Nisbett & Ross, 1980; Stanovich, 2001), (b) belief in the law of small numbers (Tversky & Kahneman, 1971), and (c) the gambler’s fallacy and hot-hand beliefs (e.g., Gilovich, Vallone, & Tversky, 1985).

Professors could illustrate estimation and the everyday use of biased samples to estimate population parameters by discussing phenomena such as the fundamental attribution error (Ross, Amabile, & Steinmetz, 1977), false consensus effect (Kulig, 2000), illusion of transparency (Gilovich, Savitsky, & Medvec, 1998), and spotlight effect (Gilovich, Medvec, & Savitsky, 2000). For the fourth concept—correlation—professors could teach students to assess simple, everyday correlations between dichotomous events (e.g., Gilovich, 1991). Common difficulties in assessing everyday correlations (e.g., the use of incomplete data or the positive-test strategy; see Alloy & Tabachnik, 1984; Kunda, 1999; Smedslund, 1963), illusory correlations (e.g., Chapman & Chapman, 1967, 1969), and stereotypes (Chun & Lee, 1999; Hamilton & Gifford, 1976) are also relevant to the concept of correlation. Finally, Kruger, Savitsky, and Gilovich (1999) described how regression toward the mean is related to a variety of everyday phenomena, such as the performance of athletes.

A number of studies suggest that students exposed to brief training in statistical reasoning or those who take a typical statistics course can learn to apply statistical concepts—namely, the LLN and regression toward the mean—to everyday situations (e.g., Fong et al., 1986; Fong & Nisbett, 1991; Lehman, Lempert, & Nisbett, 1988; Lehman & Nisbett, 1990; Lehman & Nisbett, 1990; Nisbett et al., 1983). However, researchers have yet to determine whether teaching statistics students everyday applications of a variety of statistical concepts—the approach suggested by Fong et al. (1986)—enhances their ability to apply those concepts to everyday problems. We designed a study to address this question.

A group of statistics students (the statistics reasoning group) read material and solved practice problems designed to illustrate everyday applications of statistical concepts. A
second group (the statistics control group) did not read this material. Introductory psychology students served as a comparison group. During one semester, we examined changes in the students’ ability to use statistical reasoning with everyday problems. Consistent with previous research (e.g., Fong et al., 1986), we expected that the statistics reasoning group would improve more than the other two groups and that the statistics control group would show significant, but minor, improvement. Because introductory psychology students encountered some material related to statistical reasoning (e.g., sample bias, chance events), we thought it was possible that they would also show minor improvement. Finally, we also examined changes in students’ attitudes about the relevance of statistics to everyday life.

Method

Participants

A total of 138 undergraduate students—43 men and 95 women—participated in the study. We dropped 19 students from the study because of incomplete reasoning data. The resulting sample sizes were: statistics reasoning, n = 50; statistics control, n = 45; and introduction to psychology, n = 24. These 119 students (35 men and 84 women) ranged in age from 17 to 54 (M = 21.18, SD = 6.81).

Procedure

We assigned two sections of a statistics course taught by mathematics faculty to the statistics control group and three sections to the statistics reasoning group. We balanced, as much as possible, the class times and sizes of the two groups. All instructors used the same statistics textbook (Moore, 2000). In addition, we provided students in the statistics reasoning group with a handout (organized into six sections) designed to teach statistical reasoning about everyday problems. The first author wrote these sections, and most contained about 12 single-spaced pages. The first section was a brief introduction; the other sections covered the five statistical concepts (one per section) discussed earlier. The instructors asked students to read this material outside of class and complete the four practice problems (see Appendix for sample questions). These questions were related to many of the same concepts illustrated in the practice problems discussed earlier, but they were not the same questions.

We used the value subscale of the Survey of Attitudes Toward Statistics (SATS; Schau, Stevens, Dauphinee, & Del Vecchio, 1995) to measure students’ attitudes. This subscale consists of nine items—rated on a scale ranging from 1 (strongly disagree) to 7 (strongly agree)—intended to measure attitudes about the “usefulness, relevance, and worth of statistics in personal and professional life” (pp. 869–870).

Later in the semester, we administered the SATS at the end of the semester. Most of the instructors gave students the statistical reasoning questions—two to four at a time—on exams throughout the semester. However, one statistics reasoning and one statistics control instructor administered all of the reasoning questions with the final exam. Finally, most of the instructors offered students extra credit for answering the reasoning questions; one statistics control instructor did not.

Two undergraduate students, blind to the experimental conditions, worked independently to code responses to the statistical reasoning questions and resolved disagreements through discussion. They coded a response as correct only if it reflected a correct application of the appropriate statistical concept (cf. Fong et al., 1986). To examine interrater reliability at the beginning of the study, we used a sample of coded pretest responses from 47 participants. The coders made a total of 470 judgments and agreed on 453 (96%) of them.

Results

For each participant, we calculated pretest and posttest totals for the number of statistical reasoning questions answered correctly. We also calculated each participant’s average score on the SATS value subscale for both the pretest and posttest. To determine the similarity of the groups at the beginning of the study, we performed an independent measures ANOVA

<table>
<thead>
<tr>
<th>Group</th>
<th>Pretest M</th>
<th>Pretest SD</th>
<th>Change M</th>
<th>Change SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intro to Psychology</td>
<td>1.17</td>
<td>0.76</td>
<td>0.54</td>
<td>0.93</td>
</tr>
<tr>
<td>Statistics control</td>
<td>1.29</td>
<td>1.12</td>
<td>0.44</td>
<td>1.14</td>
</tr>
<tr>
<td>Statistics reasoning</td>
<td>1.16</td>
<td>1.09</td>
<td>1.76</td>
<td>1.64</td>
</tr>
</tbody>
</table>

Note. Pretest scores for statistical reasoning could range from 0 to 10 questions correct. For attitudes toward statistics, average scores could range from 1 (strongly disagree) to 7 (strongly agree). Change = posttest–pretest.
on participants’ pretest scores (see Table 1). There were no significant differences among the groups in either statistical reasoning scores or attitudes: $F(2, 116) = 0.21, p = .82$; and $F(2, 116) = 2.42, p = .09$, respectively. Because our main interest was in examining changes in students’ statistical reasoning and attitudes toward statistics, we calculated each participant’s change score (posttest–pretest) for both measures (cf. Lehman et al., 1988; Lehman & Nisbett, 1990). As shown in Table 1, all three groups exhibited positive gains in statistical reasoning from the pretest to the posttest. An independent measures ANOVA revealed a significant effect of group on the reasoning change scores, $F(2, 116) = 13.31, p < .001$. To estimate the size of this effect, we calculated the $f$ statistic (Cohen, 1988) and found a large effect, $f = .48$.

A Tukey’s HSD test revealed that the statistics reasoning group ($M = 1.76$) exhibited a larger change in reasoning than both the statistics control ($M = 0.44$) and introductory psychology ($M = 0.54$) groups (both $ps < .001$). The latter two groups did not differ significantly, $p = .96$. Nevertheless, the gain exhibited by each group was significantly different from zero: $t_{\text{statistics reasoning}}(49) = 7.61, p < .001$; $t_{\text{statistics control}}(44) = 2.62, p = .01$; $t_{\text{intro psych}}(23) = 2.85, p < .01$. Finally, the mean attitude change scores for the three groups were not significant, $F(2, 110) = 1.90, p = .16$.

**Discussion**

Consistent with our hypothesis, the statistics reasoning group showed gains in statistical reasoning that surpassed the other two groups. In fact, their gain was four times higher than that shown by the control group of statistics students—an impressive result given the minor role of the statistical reasoning handout in the statistics course. This result supports and extends previous researchers’ (e.g., Fong et al., 1986) findings that training can improve students’ ability to use statistical reasoning with everyday problems.

Nevertheless, at the time of the posttest, the statistics reasoning group answered an average of slightly less than 30% of the questions correctly. Although this result is not unusual for this type of study (cf. Fong et al., 1986), one might question the practical significance of this finding. Our participants might have exhibited higher levels of statistical reasoning if the statistics professors had devoted more attention to this topic or if our participants had been exposed to the topic in several courses. However, we believe that increasing people’s ability to reason statistically about everyday situations—even just one or two situations—could have immense practical significance. For example, Myers (2001) explained that people’s exaggerated fear of flying in the wake of the September 11 terrorist attacks has had a dramatic impact on the airline and travel industries, and it could increase the number of drivers and traffic deaths. This problem might be lessened if people understood the role of the availability heuristic in inflating estimates of the probability of being involved in an airplane accident.

The significant gain in statistical reasoning we found in the statistics control group replicates previous research (e.g., Fong et al., 1986). However, this gain was small and no greater than that exhibited by introductory psychology students. Our results suggest that combining training in basic statistical concepts with training in everyday applications of those concepts—as we did in our statistics-reasoning group—will result in far greater gains. Fong et al. (1986) found similar results.

The lack of significant changes in the statistics students’ attitudes toward statistics is a disappointing result, but one that was not entirely unexpected in light of the mixed results found by previous researchers (e.g., Shultz & Koshino, 1998; Sorge, Schau, Hubele, & Kennedy, 2000; Waters, Martelli, Zakrajsek, & Popovich, 1988). Perhaps taking more than one statistics course would have a greater impact on students’ attitudes.

In sum, statistical concepts can be useful for making decisions in everyday life. We hope that we have given psychology instructors some useful ideas for enhancing their students’ ability to engage in statistical reasoning about everyday problems.

**References**


Appendix

Example Reasoning Questions

**Probability/Chance**

A man who wants to find out more about his future visits a psychic, hoping that she will give him some insight. She makes a number of predictions about events that will happen to him in the future, and one of them is that he will have problems with his car. Several months later, he gets into a car accident, and he has to spend a month without his car while it is being repaired. Afterward, he visits the psychic again to hear additional predictions about his future. One thing that she predicts is that he will receive good news at his workplace. A few weeks later, he learns that he is being promoted to a better job. Should he be impressed with the psychic's ability to predict the future? Explain.

**Law of Large Numbers**

Imagine that you are in charge of recruiting basketball players for a college basketball team. You have one more slot open on your team, and you are trying to decide between two players—Jason and Tom. Although the reports you received from their high school coaches indicate that Jason is a better player than Tom, you are very impressed with Tom during an hour-long tryout at your college. Tom made more baskets than did Jason, and he had more rebounds as well. Which player would you choose? Why?

**Estimation/Sample Bias**

Justin, a resident of Minnesota, often visits his friend Allen in Wisconsin. Justin loves to try different restaurants, and each time he visits Allen in Wisconsin, Allen takes him to a different restaurant for dinner. Moreover, when Allen visits Justin in Minnesota, Justin often takes him out for dinner. After several years of visiting each other’s towns, Justin says to Allen, “You know, the restaurants in your town are so much better than the restaurants in my town; I should move to Wisconsin.” Allen replies, “That’s funny, because I’ve always thought just the opposite; the restaurants in your town are better.” Why do you think Justin and Allen arrived at such different conclusions?

**Correlation**

Greg believes that there is a relation between the weather outside and the pain in his lower back. During the past several months, he noticed it was cold outside on most of the days he had a backache. Do you agree that his pain is related to the weather? Why?

**Regression Toward the Mean**

A college professor notices that students who score 100% on the first exam in her classes typically score lower than 100% on the second exam. Why do you think this happens?

Notes

1. We thank Heather Re and Karen Oswald for coding the statistical reasoning responses and for entering the data into SPSS.
2. Send correspondence to Timothy J. Lawson, Psychology Program, College of Mount St. Joseph, Cincinnati, OH 45233; e-mail: tim_lawson@mail.msj.edu.
Making Statistics Come Alive: Using Space and Students’ Bodies to Illustrate Statistical Concepts

Jane Marantz Connor
Binghamton University

This article describes several exercises that illustrate statistical concepts using students’ bodies and the physical space in the classroom. The concepts of central tendency, variability, correlation, and regression are among those illustrated. These exercises encourage both active learning and the visual-spatial representation of data and quantitative relations. Students evaluate the exercises as both interesting and useful.

Many psychologists have observed that statistics is an important course in the psychology curriculum (Bartz, 1981), but also an extremely challenging one for faculty to teach effectively (Conners, Mccown, & Roskos-Ewoldsen, 1998). Many students have negative associations of mathematics from previous encounters, few psychology students describe themselves as inherently interested in the subject matter of statistics, and many consider the subject matter “dry” or “boring” (Hembree, 1990; Tobias, 1995).

In this article I describe several exercises I have used in the undergraduate statistics course that students find interesting and useful. The exercises incorporate students’ bodies and their locations and arrangement in the classroom to convey statistical concepts. The exercises may also be helpful for students who learn best from visual or spatial representations, rather than through the exclusive use of verbal or mathematical symbols.

First Day of Class

At the first class period I explain that I find it useful at the beginning of a course to explore students’ reasons for enrolling in my class. “How can I find out why you are taking the class?” I ask. They often suggest constructing a survey. “But I would like to find out right now,” I say. Eventually, I tell them that I am going to read a list of reasons students might have for taking this course, and I ask them to indicate whether each of these reasons is, in fact, one of their reasons. However, instead of simply having them raise their hand, providing dichotomous data, I explain that I find it more useful to collect data in such a way that there is a potential range of responses. So, if they find they do not agree at all with a particular reason for taking the course, they are to keep their hands completely down at their side (they are all standing). If they agree slightly, they should raise their hands slightly; if they agree more strongly, they should raise their hands higher, with the height they raise them corresponding to the extent to which they agree with that reason. For total agreement they should raise their hands completely and move them about vigorously, indicating their enthusiastic agreement with that reason.

I read the list of reasons: “because a friend recommended this as a good course,” “because it sounded interesting,” “because I have always had a curiosity about statistics,” and “because it is a required course for the major.” Invariably, the students are quiet and inactive until I read the last reason, at which time almost all the students have their hands fully raised and are moving about. They laugh, and it seems that tension is reduced as we talk about why students often feel anxious about taking this class, what assistance is available, and related concerns. We discuss different kinds of variables in the context of this exercise: dichotomous versus multivalued, discrete versus continuous, and quantitative versus qualitative. We talk about other ways that I could have obtained the same or similar information. For example, the students could have rated their support or nonsupport for each reason by using a numerical scale, varying their facial expressions, or applauding. Each of these approaches would yield a different type of data, with somewhat different characteristics. Presumably, the basic underlying information about why the students were taking the course would be the same.

Central Tendency, Variability, and Shapes of Distributions

In a class of 50 or fewer students I conduct the following demonstration with the whole class. In a larger class I ask for 10 to 20 volunteers (approximately half men and half women).

I ask the students to visualize a number line going across the front of the room from left to right, in front of the first row of seats. At the extreme left of the number line, against the left-hand wall is the number 0, which stands for absolutely do not like at all. At the extreme right is the number 10, which stands for completely and totally like. The spot in the middle stands for 5, neither like nor dislike. I ask the volunteers to position themselves on this number line according to how they feel about the stimuli that I describe. If there is more than one person who has the same feeling about a stimulus, then one person stands on the number line at the appropriate location representing that feeling, and the others stand single file in front of that person. What emerges, of course, is a human frequency distribution or histogram.

I choose stimuli that elicit distributions that differ in shape, central tendency, and spread and that do not require the students to reveal information that would make them uncomfortable. The request “Position yourself on the number line according to how you feel about using computers” typically elicits a relatively symmetrical distribution with a center around 5. I ask students in the class to try to describe this distribution verbally, so that we might characterize how it differs from other distributions we observe. Where is the center? How spread out is the set of scores? What is the range? Do the scores cluster? I position myself on the line about where the median is to facilitate a discussion of central tendency.
When I say “Position yourself on the number line according to how you feel about chocolate” I typically get a pile-up of students at 9 or 10 and a few trailing down to 5, yielding a negatively skewed distribution. I also get an interesting pattern in response to the instruction “Position yourself on the number line according to how you feel about watching football on television.” This question frequently yields a bimodal distribution (often differentiated by gender). Usually there is no or little overlap between the male and female distributions. Sometimes, however, the responses to this question will include an outlier—a man who intensely dislikes watching football or a woman who likes it intensely.

Correlation, Scatterplot, and Regression

One of my favorite demonstrations is what I call the human scatterplot. This demonstration involves the whole class and requires some advanced planning, especially when done in a large lecture hall, but I routinely use this demonstration with as many as 200 participants. Students position themselves in a given row in the lecture hall according to their value on one variable and then move toward the left or right end of the row, according to their value on another variable.

Two variables that I have used that always seem to result in a successful demonstration are height and shoe size. The week before the demonstration I ask the students to submit a piece of paper with their height, shoe size, and gender. I use this information to determine which heights to assign to which rows and what shoe size labels to give to the endpoints of the rows. I also use it to calculate the exact equation for the regression line, which I present in the lecture and discussion that follow. Naturally, the heights and shoe sizes differ for men and women. I usually do this demonstration with the men and women separately, so that each group can both participate in the exercise and view the scatterplot as a whole, but it does work with the genders combined.

In my classroom there are 10 rows of seats. I designate each row for a particular height; for women I might use Row 1: 60.9 in. or shorter; Row 2: 61.0 in. to 61.9 in., and so on. Then I tell the women that the left-most part of each row stands for shoe size 5 and the right-most part stands for shoe size 10+. They are to position themselves in a particular row and part of the row according to this information. After a little noise and commotion we discuss the pattern that results. I ask them to note that there are no people in the front rows on the right end. Why is this? They can readily see that there are no women who are about 5 feet tall and have shoe sizes larger than 10. Similarly, the left-end rear is empty. Again, they see that women who are 6 feet tall do not have size 5 shoes. I have them notice the tendency for people to fall on a diagonal line through the classroom, indicating a positive relation between the two variables. I ask all people 64 in. tall to raise their hands and to call out their shoe sizes. The students see that there is variability in the shoe sizes of this group of people, but that there is less variability than in the class as a whole. I do the same with people whose height is 68 in. or so that the students can see both the relation between height and shoe size and the variability about the regression line. On the basis of the information in the human scatterplot I ask them to predict the shoe size for an individual of a particular height. I also ask the middle person in each row to raise his or her hand to make the regression line more visible. (An anonymous reviewer suggested that the middle persons might all hold onto a long piece of crepe paper or ribbon.) Later when I present the formal calculations using the information they provided, I show how the slope of the regression line indicates the number of minutes per week they engage in exercise or sports, the number of hours per week they watch sports on television, or the number of math courses they have had.

Other Applications

With a little practice it is not difficult to develop other applications. I have used the same basic approach to illustrate the effects of linear transformations on the central tendency and variability of scores as well as a demonstration illustrating the central limit theorem. As a side benefit, I invariably find that the physical activity associated with moving about the classroom has a positive effect on the students’ attentiveness when they return to their seats.

Evaluation and Conclusions

These demonstrations routinely receive a very positive response from the students. In an end-of-the-semester survey, a sample of 165 students evaluated these demonstrations along with several other components of the course for interest and usefulness, on a scale from 1 (not at all interesting or useful) to 7 (extremely interesting or useful). Repeated-measures ANOVAs showed that the students made distinctions among these components, $F(4, 656) = 31.8$, LSD = 0.5 for interest ratings, and $F(4, 656) = 8.3$, LSD = 1.0 for usefulness ratings. The students perceived the exercises as similar in interest value to the videos ($M = 6.7$ and 6.4, respectively) and more interesting than the text ($M = 4.9$), the lectures ($M = 5.1$), and the computer assignments ($M = 4.4$). They viewed the exercises as similar in usefulness to the text ($M = 6.5$ for exercises; $M = 6.6$ for text), the lectures ($M = 5.6$), and the videos ($M = 6.1$), but more useful than the computer assignments ($M = 4.1$).

I find that these exercises are helpful in reducing the negative associations that many students have with statistics and mathematics. They help to make the concepts concrete, which can facilitate learning (Conners et al., 1998), and they involve active learning, which improves student motivation and learning (Benjamin, 1991). They may also assist with the development of spatial intelligence (Gardner, 1993). Finally, there is no doubt that the exercises are memorable, a characteristic not typically associated with the content of statistics courses (Conners et al., 1998). Years after students have
taken my course and forgotten the phrase “regression line,” they have described the scatterplot exercise to me, how the students were clustered along the diagonal of the room, and the relation between height and shoe size that was reflected in the fact that there were no short students with very big feet or tall students with very small feet.

References


Note

Send correspondence to Jane Marantz Connor, Department of Psychology, Binghamton University, Binghamton, NY 13902–6000; e-mail: jconnor@binghamton.edu.

Helping Students Gain Insight Into Mental Set

Richard A. Griggs
University of Florida

Difficulty in problem solving is an important topic in the thought-language chapter in introductory textbooks. Textbook authors typically use water-jar problems to illustrate mental set, one of the major barriers to successful problem solving. Such problems, however, only exemplify mental set created within the actual problem setting. To extend such discussions, I describe an engaging class activity that uses 4 related series problems to illustrate mental set and its more general negative effect on problem solving. The nature of the interfering mental set therefore varies across the four problems. Thus, these problems both exemplify the general nature of mental set and its key role in making problems more difficult and set the stage for a discussion of the more general concept of mindlessness.

Activity Materials and Procedure

For a better understanding of this activity, you should attempt to solve the four series problems given in Table 1 before reading further. One can present the four problems on a chalkboard, overhead transparency, or Microsoft PowerPoint slide. Present each problem one at a time and in the order given in the table. As you proceed from one problem to the next, leave all prior problems in view and do not give any of the answers until you have presented all of the problems.

Ask students who have seen a problem before or solve a problem not to give the answer, but rather to think of clues to help their classmates overcome mental set and gain insight into the problem’s solution. These students can also testify that the problems all have rather simple answers, which keeps other students motivated to search for solutions. In addition, the students providing clues gain a better understanding of mental set and its negative effect on problem solving via their attempts to find facilitating clues. The clues inevitably lead to some in-class “Aha!” insight experiences that students cannot suppress. In addition to student clues, provide...
Table 2. Evaluative Data for the Applied Psychology Exercise

<table>
<thead>
<tr>
<th>Item</th>
<th>Fall ¹</th>
<th>Spring ²</th>
<th>Fall ³</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>1. Interesting</td>
<td>4.27</td>
<td>.70</td>
<td>4.36</td>
</tr>
<tr>
<td>2. Use in future classes</td>
<td>4.40</td>
<td>.74</td>
<td>4.55</td>
</tr>
<tr>
<td>3. Understand people and forces that shaped applied psychology</td>
<td>4.40</td>
<td>.51</td>
<td>4.45</td>
</tr>
</tbody>
</table>

Note. Participants used a scale ranging from 1 (strongly disagree) to 5 (strongly agree). ¹N = 16. ²N = 11. ³N = 10.

To assess the effectiveness of this exercise, students rated its interest value, whether it should be used in future classes, and the extent to which it helped them understand the people and events behind the history of applied psychology. I used the same 5-point rating scale used in the James exercise. Results for all three semesters appear in Table 2.

Conclusions

From my perspective as instructor, both activities worked exceedingly well. Each exercise used interactive discussions and had a writing-to-learn component (Fulwiler & Young, 1990). This format allowed all students to make a unique contribution to the issues under consideration. Giving voice to students’ ideas about the history of psychology also enhanced their interest in and understanding of the material. Rather than processing information in a passive and shallow manner, students had to weigh carefully a number of important considerations for making judgments about historical figures. Given the ratings students assigned to these exercises, I believe that they appreciated that opportunity. Of course, I will follow the advice of my students and use the exercises again. I look forward to future votes on candidate James and hope to continue to give new meaning to dating in a history of psychology course.

References


Notes

1. I thank the anonymous reviewers for their helpful comments on previous drafts of this article.
2. Send correspondence to David Zehr, Department of Psychology, Plymouth State University, Plymouth, NH 03264; e-mail: zehr@mail.plymouth.edu.

Cinema and Multiple Regression

Carmen García
Department of Social Psychology and Methodology
Autónoma University of Madrid

Maria Teresa Ruiz García
Department of Biological and Health Psychology
Autónoma University of Madrid

We presented a task to increase students’ involvement in learning multiple linear regression. Fifteen students predicted how they would rate a film they had not seen, basing their criteria on several film critics’ reviews of a series of movies. The evaluation of the exercise showed that it helped students understand multiple linear regression as well as in increasing their interest in this technique.

To teach multiple linear regression, we had students predict how they would rate a film they were going to see for the first time. The predictions were based on a regression equation generated from reviews of well-known film critics on a number of films the students had earlier seen and rated.

This activity attempted to achieve two aims. On one hand, studies on social communication show that the personal relevance a message has for its recipients is a good pre-
dictor of the attention and elaboration that the messages receive (García, Ponsoda, & Estebananz, 2000; Zaichkowsky, 1985). Based on this idea, we gave students data that were both easily obtained and directly related to their daily lives, so they would find their analysis personally relevant. The assessment of the films previously seen by the students provided them with data appropriate for a meaningful regression.

The second aim of our activity was to reduce the anxiety traditionally associated with statistics as a subject (G. A. Allen, 1981). The use of data related to the students’ daily lives and the relation of statistics to other materials contributed to achieving this goal.

Method

Participants

Fifteen first-year doctoral students in psychology and education took part in the activity, which took place in a multivariate techniques course. Although all students had some basic knowledge of descriptive and inferential statistics, they had none about multivariate techniques.

Data Collection

The students gathered reviews of a wide range of films they had recently seen. We selected reviews from those written by film critics published in eight national newspapers and specialized film magazines.1 Each student had to construct a data matrix made up of (a) several films that had been shown at the cinema, (b) the ratings the critics gave the films, and (c) their own ratings, using the same scale as the critics (i.e., 1 = bad, 2 = fair, 3 = normal, 4 = good, 5 = excellent). Table 1 shows an example of a data matrix.

The Linear Regression Model

We explained that the basic concepts used in any linear regression model applied to the ratings each student had placed in his or her data matrix: the predictors (the ratings of the film critics), the criterion variable (how the students rated the films they had seen), the forecast (the predicted rating, using the regression equation), and the residuals (the difference between the ratings the students gave and those predicted by the equation). We also took note of the three main sources of the forecast error: (a) errors of a general character in the measurement of the variables (e.g., errors that arose when constructing the data matrix), (b) the absence of relevant predictors, and (c) incorrect choice of the functional form of optimum prediction for the data (nonlinear relations between predictors and the forecast variable).

Before asking the students to perform the regression, we explained the steps of the experiment and instructed the students to check that their data met the requisites. In this sense, we explained that the data they were going to use did not meet regression assumptions. The data were clearly ordinal, whereas regression assumes interval or ratio scale data. Although the data did not meet the assumptions, the predictions the students made were fairly accurate.

We informed the students that the number of predictors (film critics) had to be fewer than the number of films to avoid instability in the coefficients of regression. The number of movies they rated was variable, but never less than 40. We also commented on the lack of perfect linear regressions in the critics’ ratings.

In constructing the regression model the students used SPSS. Before generating the equation, we showed the different methods for choosing predictors. We introduced the difference between theoretical and statistically guided methods. The students decided which variables to include or exclude from the equation in the inclusion and elimination methods. The criterion was exclusively theoretical. For example, an individual could decide to construct an equation with the reviews of just two critics because they write in the only newspaper the individual usually reads.

In our activity we worked with statistically guided criteria methods (forward, backward, and stepwise). Each student generated three equations, each using one of the methods. They compared the determination coefficients of each model and the chosen predictors to know which critics could best predict their rating for the films. The students chose the most adequate model, which they then used to predict the ratings of some films they were going to see. Table 2 shows the regression equations obtained by some of the students, with their respective goodness-of-fit statistics.

Each student could observe that the different methods produced different functions along with different values of the goodness-of-fit statistic. They also compared the multiple correlations to know if the differences were statistically significant and to discover which of the functions maximized the accuracy of prediction. In the example shown in Table 2, the students chose the third of the equations for their forecast because it explained a significantly higher percentage of variance, $F(2, 13) = 6.23, p < .05$.

The time required for the exercise was approximately 8 hr. This time included the explanation of the technique, data collection, statistical analysis, interpretation of the results, and the assessment.

Assessment

At the end of the course, we asked students to anonymously evaluate this approach by responding to an open-ended question about the usefulness of the task. The assessment data collected showed that the exercise helped a great deal in the students’ understanding of multiple linear regression. The best evidence was that the activity actually produced the beneficial results we had expected (i.e., the involvement of the students). Many of them used the equations they obtained, or constructed new ones for their friends, to predict which among the several films they should see when...
The result is consistent with the predictions that Mathie et al. (1993) made for practices that involve active learning. Some of the written comments by the students were: “I found the exercise on the films useful in a nice and pleasant way” and “The classes deserve good marks because they were pleasant and enjoyable.” Other opinions were similar: “The work I had to do on this technique has been fundamental, … it helped me to understand its possibilities and to think about them on application” and “I had to deal with some aspects I thought I knew but had not really assimilated. I now have the feeling I mastered those techniques.”

Conclusions

In this article, we have presented a method with the aim of achieving a better prediction of a variable starting from other variables. The greatest advantages of our task were that (a) it allowed the involvement of the students in the data collection, with little time and effort needed to obtain sufficient data; and (b) it allowed the students to be interested in the predictions they achieved. Moreover, according to the students’ personal interests, the same exercise could be carried out with the use of other data sets, on the ratings of restaurants, cafes, hotels, books, songs, video games, and online video games, for example.

References


Notes

1. An early version of this article was presented at the First National Conference on Teaching of Psychology, Valencia, Spain, February 1999.

2. We thank the editor and reviewers for their valuable suggestions for revisions to an earlier draft of this article.

3. Send correspondence to Carmen García, Department of Social Psychology and Methodology, Autónoma University of Madrid, 28049 Madrid, Spain; e-mail: carmen.garcia@uam.es.

Table 1. Data Matrix Used in the Multiple Linear Regression

<table>
<thead>
<tr>
<th>Movie</th>
<th>El Mundo</th>
<th>El País</th>
<th>Fotogramas</th>
<th>RN1</th>
<th>ABC</th>
<th>Diario 16</th>
<th>Dirigido por</th>
<th>Ser</th>
<th>Me</th>
</tr>
</thead>
<tbody>
<tr>
<td>Everyone Says I Love You</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>The English Patient</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Secrets of the Heart</td>
<td></td>
<td></td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

Note. The values represent the ratings the critics and the students gave the films on a 5-point scale ranging from 1 (bad), 2 (fair), 3 (normal), 4 (good), to 5 (excellent).

Table 2. Regression Functions Obtained by the Forward, Backward, and Stepwise Methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Function</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stepwise</td>
<td>$Y' = .67 + X_1$</td>
<td>.61*</td>
</tr>
<tr>
<td>Forward</td>
<td>$Y' = .67 + X_1$</td>
<td>.61*</td>
</tr>
<tr>
<td>Backward</td>
<td>$Y' = 1.13 + 1.29X_1 + .67X_2 - .89X_3$</td>
<td>.79*</td>
</tr>
</tbody>
</table>

Note. $N = 15$. Of the eight predictors that could form part of the equation, we chose only three ($X_1, X_2, X_3$); they corresponded to the ratings of the critics from ABC, Dirigido por, and Radio Nacional 1. *$p < .001$.
A Hands-On Exercise Improves Understanding of the Standard Error of the Mean
Robert S. Ryan
Teaching of Psychology 2006 33: 180
DOI: 10.1207/s15328023top3303_5

The online version of this article can be found at:
http://top.sagepub.com/content/33/3/180

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Society for the Teaching of Psychology

Additional services and information for Teaching of Psychology can be found at:

Email Alerts: http://top.sagepub.com/cgi/alerts

Subscriptions: http://top.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Jul 1, 2006

What is This?
A Hands-On Exercise Improves Understanding of the Standard Error of the Mean

Robert S. Ryan  
Kutztown University

One of the most difficult concepts for statistics students is the standard error of the mean. To improve understanding of this concept, a group of students used a hands-on procedure to sample from small populations representing either a true or false null hypothesis. The distribution of 120 sample means ($n = 3$) from each population had standard errors that closely approximated those of the theoretical sampling distributions, thereby illustrating how the Central Limit Theorem provides a standard error to use for hypothesis testing. Performance on an exam about the standard error of the mean was significantly better for the students who had completed this exercise than for students in a control group.

Psychology students generally consider inferential statistics to be one of the most challenging subjects in their curriculum. The logic of statistical inference depends on understanding the standard error and the related concepts of hypothetical sampling distributions under either a true or a false null hypothesis. Several studies have described hands-on exercises designed to facilitate learning about sampling distributions (e.g., Dyck & Gee, 1998; Gourgey, 2000; Johnson, 1986; Rossman & Chance, 2000). However, these studies did not address whether understanding the standard error, in particular, can be improved specifically by hands-on experience with constructing both a null-true and a null-false sampling distribution.

The Dyck and Gee (1998) and Johnson (1986) articles presented exercises similar to the one in this study, but with some important differences. In both studies, the participants themselves did not randomly draw members from a population to form samples. In Dyck and Gee’s study, each participant determined the value for one member of a population by counting the number of blue M&Ms® in a package, and the instructor selected members from that population to form the samples. In Johnson’s study, the instructor presented each student with three or four equal-sized samples that he told them had been randomly selected from a population. In addition, in both studies participants did not draw samples from a population that could have had either a mean specified by a null hypothesis that was true or some other mean. In Johnson’s study, the participants knew the population mean in advance. In Dyck and Gee’s study, the participants knew that all of the samples came from a population with the same mean, which they calculated after the sampling process.

These procedures are appropriate for exercises that focus on improving understanding of the shape of the sampling distribution and the relation between its mean and the population mean. However, a critical idea for understanding the standard error is the distinction between the variability of members of a sample versus the variability of the sample means. An exercise in which participants draw the members of the samples themselves, as well as place the means of the samples in a distribution, would provide a way for them to experience the critical distinction directly. Furthermore, drawing such samples without knowing from which of two populations they came would enable the participants to notice that it is specifically the variability of sample means around the mean of all the sample means that is most directly relevant to deciding whether to accept or reject the null hypothesis. Therefore, this article presents a hands-on exercise in which students directly experienced the hands-on construction of both a null-true and a null-false sampling distribution.

The reason for the usefulness of such an exercise stems from a common problem that often interferes with students’ understanding of the standard error, namely that they do not clearly distinguish between distributions of scores and a distribution of sample means. Most texts on statistics for the behavioral sciences describe a sampling distribution of the mean as consisting of the means of all possible samples of some size taken from the same population (e.g., Christensen & Stoup, 1991; Gravetter & Wallnau, 2000; Grimm, 1993; Hurlburt, 2003; Witte & Witte, 2004). Therefore, understanding the idea of a sampling distribution requires that students understand that they must now consider one of the measures that they formerly calculated for a set of scores (i.e., the mean) to be one of many single scores in yet another distribution (i.e., the sampling distribution). At the same time students are trying to learn this distinction, however, they are already trying to deal with a great deal of abstractness. For example, the characteristics of a sampling distribution, such as its central tendency and...
variability, are abstract ideas that students measure using mathematical tools, which are also abstract. Furthermore, the measurements apply not to something concrete (e.g., measuring the area of a room), but to a distribution of scores, which is itself an abstract concept. Therefore, students may focus on the idea of “every possible sample” as the most salient concept in the definition of a sampling distribution. As a result, they may believe that the sampling distribution is simply an amalgamation of a large number of samples, which would lead to the misconception that the standard error measures the variability of individual scores rather than the variability of sample means.

Misconceptions are often difficult to overcome specifically because people disregard correct information about the misconception (Eaton, Anderson, & Smith, 1984). However, information that students generate for themselves is harder to disregard (Slamecka & Graf, 1978). In fact, Chi, deLeeuw, Chiu, and LaVancher (1994) found that self-generated information was specifically beneficial for the especially difficult task of overcoming misconceptions.

Overcoming students’ common misconception about the standard error rests at least in part on distinguishing between a distribution of scores and a sampling distribution. Therefore, this exercise maximizes the likelihood that students would notice that the distribution (of which the standard error measures the variability) consists of sample means, not individual scores.

A new aspect of the exercise that was not present in previous hands-on constructions of sampling distributions highlighted this distinction. Specifically, students constructed sampling distributions by drawing samples both from a null-hypothesis-true and a null-hypothesis-false population. This innovation focused students’ attention on how a researcher is able to draw an inference about a population mean from a sample mean, even without knowing from what population the sample actually came, specifically because the sample mean is one of many sample means in a hypothetical distribution.

Method

Participants

The participants in the control group were 17 graduate students in an inferential statistics course in a counseling psychology master’s program at Kutztown University in the fall of 2001. The participants in the experimental group were 29 graduate students in the same course in the fall of 2002.

Materials

Populations for the exercise. I used two sets of bags to hold slips of papers showing the values in the populations representing the true null hypothesis and the alternative hypothesis. Figure 1 shows the values that I used and how many of each value were in the bags. Both populations consisted of the same values but were skewed to produce different means. As a result, the participants could select samples from populations with different means, but could not tell from which population they were selecting by looking at the values.

Theoretical sampling distributions. As shown in Table 1, the difference between the population means, and therefore between the theoretical sampling distribution means, was 2.4. Both sampling distributions consisted of the means of all 125 possible samples ($n = 3$) using sampling with replacement.

Exam. Both groups took an exam on sampling distributions. A subset of 15 items that were identical for the two groups assessed the effectiveness of the exercise. Where there were multiple items on the same type of question, intercorrelations assessed reliability.

Evaluation survey. The students evaluated the hands-on exercise by taking a short survey. It consisted of one general open-ended question and five more specific questions.
Procedure

I introduced the exercise with an example to remind the students that researchers take a random sample to make an inference about a population mean. I told the students that there were six bags that had a population in which the null hypothesis that \( \mu \leq 1.8 \) was true and another six that had populations in which the alternate hypothesis that \( \mu > 1.8 \) was true. I instructed the students to randomly select samples of \( n = 3 \) using sampling with replacement.

After the students selected each sample, I told them to calculate its mean. If the sample mean was 4 or greater, the students rejected the null hypothesis. Otherwise, they failed to reject. Although this criterion resulted in the unusual significance level of .128, it facilitated calculating the empirical Type I error rate because every sample mean would fall clearly above or below the criterion. Finally, they recorded the value for each member of the sample, the sample mean, and the decision. Twelve small groups of students selected 10 samples, guessed which population their samples came from (in all cases the guesses were correct), and then traded bags with a group that had the opposite population and selected 10 more. I used the groups’ recorded results to form empirical sampling distributions that the students compared to the theoretical sampling distributions in the next class. The entire exercise took about 30 min.

In the next meeting of this class, I illustrated the Central Limit Theorem by presenting the empirical sampling distributions formed from the 120 randomly selected samples from each population. Table 1 shows that the means, standard errors, Type I error rate, and power of the empirical sampling distributions were very close to those for the theoretical sampling distributions. I taught the concepts of sampling, sampling distributions, the Central Limit Theorem, errors of inference, and power to both groups in the same way except that I referred the hands-on group to the sampling distributions they had constructed, whereas I referred the control group to an example that I constructed at the blackboard and to textbook material. Due to the greater length of time required for explaining and conducting the exercise, the hands-on group spent two classes accomplishing the same lessons that the control group accomplished in one class. Both groups took their exam on sampling distributions at the beginning of the class that immediately followed the class (or classes) on sampling distributions.

Table 1. Characteristics of the Two Types of Sampling Distributions of the Mean Under Different Hypotheses

<table>
<thead>
<tr>
<th>Null Hypothesis</th>
<th>Sampling Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Theoretical</td>
</tr>
<tr>
<td>True ( M )</td>
<td>1.80</td>
</tr>
<tr>
<td>Standard error</td>
<td>1.39</td>
</tr>
<tr>
<td>Type I error rate</td>
<td>0.13</td>
</tr>
<tr>
<td>False ( M )</td>
<td>4.20</td>
</tr>
<tr>
<td>Standard error</td>
<td>1.39</td>
</tr>
<tr>
<td>Power</td>
<td>0.72</td>
</tr>
</tbody>
</table>

Results

Exam Results

As predicted, the hands-on group achieved higher exam scores (\( M = 94.7, SD = 6.5, N = 29 \)) than the control group (\( M = 90.2, SD = 10.2, N = 17 \)). The scores of the hands-on group ranged from 80% to 100% correct, whereas the scores of the control group ranged from 63.3% to 100% correct. Because the sample sizes for the exam scores were different and the variances were unequal, \( F(16, 28) = 2.47, p = .018 \), I first analyzed the data with an unequal variance \( t \) test. Although this test did not reveal a significant difference, \( t(23.7) = 1.64, p = .058 \) (one tailed), I also conducted a more sensitive test and one that used equal sample sizes. This test used the percentage of students who were correct on each exam item as the unit of analysis. In this analysis, the test items formed matched pairs of percentage correct for the two groups. Because this paired \( t \) test took into account the differences in performance due to the different test items, it showed that the hands-on group performed significantly better than the control group, \( t(14) = 3.14, p = .004 \) (one tailed). In contrast to the exam on sampling distributions, even when all the items were identical, the hands-on group did not perform significantly better than the control group on any other exam (all \( ps > .05 \)).

The reliability intercorrelations of four items that tested the application of knowledge of the sampling distribution of the mean were all significant at the .0001 alpha level. They ranged from \( r(46) = .56 \) to \( r(46) = .99 \) with an average of \( r(46) = .81 \).

Survey Results

Seventeen students (81%) reported that they believed they understood sampling distributions better as a result of doing the exercise and that the exercise was either enjoyable, fun, or both. Nine students reported other specific ideas that they understood better as a result of the exercise, including six concepts that students also have difficulty understanding (e.g., the implication of repeated sampling, the normal curve, the null hypothesis). The students rated the educational value of the exercise on a 7-point scale, ranging from 1 (not valuable at all) to 7 (very valuable). The mean rating was 5.29 (\( SD = 1.52 \)). However, because there was a positive correlation between judgment of educational value and the extent
to which the exercise was fun, $r(21) = .75, p < .005$, these opinions may have simply reflected how much the students enjoyed the exercise.

**Discussion**

The students who participated in the hands-on sampling exercise performed better on an exam about the sampling distribution of the mean than the students who had not participated, even though they performed no better on other exams. The hands-on students’ evaluations of the exercise suggested that they enjoyed it and believed that it was educationally valuable. However, given that the exercise for the hands-on group was spread over two classes, the possibility exists that their superior performance was due either to having more time to learn the material or to having spaced learning sessions.

The results for individual exam questions provided some insights, as well as some questions, about the exact nature of the benefits of the exercise. One of the exam questions asked what the standard error measures. The correct alternative stated, “How much the sample means differ from the population mean.” Selecting that alternative suggests understanding that the sampling distribution consists of sample means and that their grand mean equals the population mean. However, students often selected the incorrect alternative, “The variability of the scores in the sample around the sample mean.” Selecting that alternative shows confusion between the standard error and the standard deviation of the sample. In this study, the single greatest improvement was on that question, with only 53% of the control participants but 69% of the hands-on participants answering it correctly. This difference suggests that selecting samples and calculating their means called attention to the fact that not only is there variability among the scores in the sample, but also that the means of each of the samples differ from one another.

The other exam questions on which the hands-on group outperformed the control group, although not significantly, were those that asked about population variability and sample size. It is so important for students to understand these relations, it might be worth speculating that the failure of this result to reach significance may have been due to Type II error, even though it could also have been due to chance alone. Perhaps the reason students typically fail to understand how population variability and sample size affect the size of the standard error when they hear about these effects in a lecture is because they lack the prerequisite knowledge. If so, then perhaps the hands-on exercise provided at least some of that knowledge, although not enough so that its effect was detected at a significant level in the data reported here. Therefore, future research might profitably examine ways of modifying the exercise to strengthen its ability to improve this particular aspect of students’ understanding.

The overall exam improvement was only 4.5%, but it was an improvement over the already respectable 90.2% obtained by the control group. Furthermore, the hands-on group’s score of 94.7% was the highest score for either group on any of the exams. Because graduate students generally have more background knowledge, more motivation, or both, than undergraduate students, undergraduates may benefit even more from this exercise. Future research should address this issue.

**References**


**Notes**

1. More detailed information about any of the materials, such as the evaluation survey and all of the exams, is available by contacting the author.
2. I thank Kathleen Kleissler, Adrienne Lee, Anita Meehan, Bob Voytas, Carole Wells, and two anonymous reviewers for helpful comments on earlier versions of the article.
3. Send correspondence to Robert S. Ryan, Department of Psychology, Box 730, Kutztown University, Kutztown, PA 19530; e-mail: rryan@kutztown.edu.
We examined treatment of outliers in 40 introductory statistics textbooks for behavioral science. The majority of texts (75%) included treatment of outliers. Other than agreement in defining outlier as an extreme data point, there was great diversity in treatment. Of the 30 texts including outliers, 11 presented both the univariate and bivariate cases, 15 treated only the univariate case, and 4 only the bivariate case. The texts included 7 different operational definitions of outliers; 16 texts provided no operational definition. A supplementary analysis of outliers in 3 statistical software packages showed more consistency, although not complete agreement. We offer recommendations for treatment of outliers in introductory statistics courses for treatment of behavioral science.

Following Tukey’s (1970, 1977) pioneering work on exploratory data analysis, the concept of an outlier has become well established in the statistical literature. Research articles often refer to the occurrence and disposition of outliers in data analysis. Most introductory statistics textbooks for the behavioral sciences, as shown in this study, include reference to outliers. In addition, widely used statistical software packages include procedures for identifying outliers.

Many topics in the introductory statistics course (e.g., standard deviations, Pearson correlations) have standardized definitions; textbooks differ mainly in pedagogical techniques (e.g., number and types of examples used) rather than in substance. Formulas for calculating most routine statistics are highly standardized in different textbooks. By contrast, beyond the common definition of an outlier as an aberrant or very unusual data point, we have noticed considerable variety in how textbooks treat this concept. We investigated how introductory textbooks on statistics in the behavioral sciences cover outliers. We supplemented the textbook analysis by analyzing outliers in three statistical software packages.

For purposes of reference, we constructed Figure 1 showing the principal terms used in Tukey’s (1970) original treatment of outliers. The figure applies Tukey’s terms to a normal distribution of T scores (μ = 50, σ = 10). To keep the figure size manageable, we included only even-numbered scores. Where important reference values in T scores or σ units occur at odd-numbered scores, we placed them at the nearest even-numbered score. The crucial parts of Tukey’s scheme are (a) use of the interquartile range (IQR) as the basic unit (rather than σ units) for measuring distances; (b) a multiplier of 1.5 applied to the IQR; and (c) a two-tiered system, going 1.5 × IQR, then 3 × IQR beyond the first and third quartiles. Tukey’s original presentations did not include the term outlier. Rather, he labeled data points in the first tier as outside values and data points in the second tier as far out values. Today, use of the term outlier is nearly universal.

**Method**

We examined 40 introductory statistics textbooks (see Appendix) using several inclusion criteria. The text had to be aimed at the social and behavioral sciences; we did not include mathematical statistics books. The text had to be intro-
ductory, that is, designed primarily for the first course in statistics. The text had to be aimed at a typical three- or four-credit college course. We excluded texts intended as supplementary readings or for short courses. We included only books with copyrights within the past 10 years and published in the United States. The median year of publication for the textbooks was 2001. Our selection procedures were similar to those used by Guttmannova, Shields, and Caruso (2005); Landrum (2005); and Nevid and Forlenza (2005).

We first identified outliers in a book’s index. Based on references in the index, we examined the book’s coverage of outliers. We coded the topic on three questions: Was the topic covered at all (yes or no)? Did the book treat the univariate case (an outlier within a single distribution of scores), the bivariate case (usually represented by a scatter-gram), or both? Assuming treatment of the topic, did the book provide any operational definition, for example, a formula for identifying outliers (if yes, what formula was provided)? We also recorded some incidental information, such as alternative terms for outlier. We also analyzed treatment of outliers in three commonly used statistical software packages: SPSS, SAS, and Minitab.

**Results**

Among the 40 textbooks, 10 (25%) did not mention outliers and 30 (75%), a clear majority, included some description of outliers. Among the 30 books that did include the topic, 15 referred to outliers only in the univariate case, 4 referred to outliers only in the bivariate case, and 11 referred to outliers in both the univariate and bivariate cases.

Among the 30 books that covered outliers, only 14 provided a specific operational definition. The other 16 books defined outliers as extreme, unusual, or aberrant data points, without providing a specific criterion. In the books including an operational definition, 10 used a definition based on $1.5 \times \text{IQR}$; among these cases, 8 used only the $1.5 \times \text{IQR}$ definition, whereas 2 used that definition plus another definition. We encountered 6 other definitions: 2 using some other multiple of the IQR, etc. Examinations of outliers in three commonly used statistical software packages: SPSS, SAS, and Minitab.
Multiple of the IQR (3 × IQR beyond quartiles and 2 × IQR from median), 3 using a multiple of the standard deviation (2σ, 2.5σ, and 3σ), and 1 using percentages (top and bottom 2.5%). These different criteria will yield varying identification of outliers, in some cases substantially different results. One book explicitly stated that the criterion for identifying an outlier was arbitrary. None of the books used 2.7σ, which corresponds to the 1.5 × IQR definition in a perfectly normal distribution.

A few textbooks also formalized a distinction between subcategories of outliers. The distinctions generally corresponded with Tukey’s (1970, 1977) distinction between outside and far out values (see Figure 1). However, the clear majority of books used a single category of outliers. We encountered a few books that treated the concept of outlier but gave it some other name (e.g., outlayer or outlying scores). An earlier edition of one of the books used the term outlier. Generally, the texts that covered the bivariate case did not mention any criteria for defining outliers in such cases (e.g., the resistant line of fit; see Chatterjee & Yilmaz, 1992; Emerson & Hoaglin, 1983; Hartwig & Dearing, 1979). Two books did mention criteria for defining outliers in the bivariate case, one using standardized residuals and the other using changes in the correlation coefficient when a case was eliminated.

Among the statistical packages, SPSS (2003) contained several references to outliers. The Explore routine listed the five highest and five lowest values, which are not necessarily outliers in the usual sense. The Boxplot routine used the two-category system based on different multiples of the IQR: 1.5 × IQR for outliers and 3 × IQR for extreme cases, conforming to Tukey’s (1977) original scheme, although with different labels. The SPSS Cluster Analysis routine made some provision for identifying outliers in the creation of clusters. SPSS also referenced outliers under the Linear Regression routine, but these references took the user to plots of residuals, without providing criteria for identifying outliers. Like SPSS, SAS (2003) used the two-category system (1.5 × IQR and 3 × IQR) for the univariate case. In addition, SAS had several procedures for treating outliers in the bivariate (or multivariate) case. Minitab (2000), under the Boxplot routine, used the 1.5 × IQR formula to define potential outliers and 3 × IQR to define outliers. Thus, although Minitab used the two-tier system, its terminology (potential outliers and outliers) is unusual. We did not find any Minitab procedure for outliers in the bivariate case.

**Discussion**

Our principal purpose was to describe how textbooks treat outliers, not to critique any particular book’s treatment of the topic. We were most struck by the variations in discussion of outliers, both in terms of the types of cases treated and in terms of operational definitions. Approximately one third of the books covered only the univariate case and another one third covered both the univariate and bivariate cases; thus, approximately two thirds of the books covered at least the univariate case. The four examples covering only the bivariate case were striking because the (X, Y) point may be a bivariate outlier, whereas neither the X nor the Y data values may be an outlier within their univariate distributions. That is, a student could not possibly infer the meaning of a univariate outlier from the presentation of the bivariate case.

Perhaps most startling was the small number of books (14 of 40) that informed students exactly how to identify an outlier when working with a data set. Simply to suggest spotting unusual cases leaves students in an awkward position when analyzing data. How will students determine exactly what an unusual case is? Further complicating this issue is the diversity in the criteria presented for identifying outliers. The 1.5 × IQR definition was the favorite, although it appeared in only 25% of the books. The book using the 3 × IQR criterion for outliers appeared to be following Minitab usage, as described earlier; that is, an outlier is a value more than 3 × IQR beyond the quartiles, whereas values only 1.5 × IQR beyond the quartiles are potential outliers. This terminology is unusual, and at variance with most other sources. The 2 × IQR from median definition, used in only one book, corresponds to the 1.5 × IQR beyond quartiles definition but only for a perfectly symmetrical distribution. This seems confounding and at variance with many other sources. The 2 × IQR from median definition in comparison with the more customary definition of 1.5 × IQR beyond the quartiles. Similarly, all definitions based on the standard deviation present difficulties for skewed distributions. In a skewed distribution, application of a σ-based criterion will result in different levels of stringency for the two tails of the distribution (i.e., identifying more cases as outliers in the long tail of the distributions. In addition, use of any σ-based criterion requires recalculating the standard deviation after elimination of cases, something not required for the IQR-based criteria, except in very unusual circumstances. In general, books using a σ-based criterion did not mention either of these difficulties.

**Recommendations**

The following recommendations seem reasonable for treatment of outliers in the introductory statistics course in the social and behavioral sciences. The recommendations represent a sort of modal best practices list based on coverage in our 40 books, although there was much variation. We address the recommendations to instructors but, with minor rewording, they also apply to textbook authors. (a) The concept of an outlier should appear in any introductory statistics course for the behavioral sciences, specifically because the concept appears routinely in research reports. (b) The univariate case deserves an explicit, operational definition. The criterion 1.5 × IQR beyond the quartiles seems to be the “industry standard” and, therefore, we recommend its presentation. If presenting a σ-based criterion, the instructor should alert students to its asymmetrical conse-
quences for skewed distributions as well as the need to re-calculate the index after elimination of outliers. (c) The disposition of an outlier, after it is identified, is a separate matter, calling for prudent judgment, not the simple application of a formula. (d) We recommend presentation of outliers for both univariate and bivariate cases. The chapter on measures of variability seems the natural place for the univariate case and the chapter(s) on correlation for the bivariate case. (e) For the introductory course, we recommend against reference to the resistant line, standardized residuals, or other such procedures for the bivariate case; these topics are too advanced for the introductory course. However, the instructor may note that more advanced books do describe procedures for identifying bivariate outliers. (f) Finally, for courses employing software packages, we recommend that instructors alert students to Minitab’s (2000) unusual use of the term potential outliers as possibly confusing and to SPSS’s (2003) cross-referencing outliers to its listing of the five highest and five lowest scores because these scores are not necessarily nor frequently outliers.

References


Appendix

List of Textbooks Examined


Instructors of undergraduate research methods can introduce diverse perspectives into their courses through expanding learning units on research ethics to include extensive discussions on the responsibilities of the researcher. I provide suggestions for teaching strategies that promote multiculturalism while avoiding a deficit research perspective.

Content courses in psychology represent a natural opportunity to challenge bias, introduce diverse perspectives, and integrate cultural understandings while teaching principles of psychological knowledge (Gloria, Rieckmann, & Rush, 2000; Goldstein, 1997; Kowalski, 2000; Simoni, 1996). In this article, however, I address the need and the opportunity for psychology instructors to similarly incorporate discussions of diversity into undergraduate research methods classes. I have developed strategies to integrate diverse perspectives in both experimental and nonexperimental research classes, largely by revising the way I teach ethical principles for research in psychology (American Psychological Association, 2002). Including such perspectives is especially important because many psychology graduates are likely to work with disadvantaged communities and in applied settings where diversity arguably matters more than in research settings.

Students are often unfamiliar with the idea that research can be harmful to people or communities. Therefore, I introduce the topic of research ethics by describing studies, such as the Tuskegee Syphilis Study (Jones, 1981), that illustrate the dangers of the attitudes of majority-group researchers toward minority-group research participants and of the consequences of a lack of informed consent. My students hold small group discussions about whether the social beliefs of the researchers reflected racism and colonialism (Tuhiwai Smith, 2002, effectively drew the analogy of researcher to colonizer). Once they understand the way that research can cause harm, they are able to discuss important questions, such as these: Who pays the price for research, and who benefits from it? In what ways are social relationships reproduced in the research relationship? I also use this case to illustrate long-term consequences of such research. For example, recent survey findings reveal that over half of African Americans believe that a cure for HIV/AIDS is being deliberately withheld from sufferers (Bogart & Thorburn, 2005), perhaps because of lingering mistrust of health professionals and researchers.

I use a similar strategy of presenting a historical example, followed by small group discussion, to illustrate the way that research practices have affected indigenous peoples. Many Native communities no longer allow outside researchers to collect data (Kral, Burkhardt, & Kidd, 2002). For instance, Cree leaders in northern Quebec declared a moratorium on psychological research due to “the researchers’ refusal to accept Cree authority and to the little perceived benefits of this research for the community” (Darou, Kurtness, & Hum, 2000, p. 44). An Alaskan native community, harmed by the negative publicity generated by an earlier study on alcoholism, instituted a tribal ban on further inquiries into this important social problem (Norton & Manson, 1996). These examples offer me the chance to introduce the idea of deprivation or deficit research (Parham, White, & Ajamu, 2000), so called because it is based on the presumption that a minority community is of interest only to the degree that it poses research problems for majority-group researchers. Deficit research socially constructs communities as being naturally or culturally deficient relative to the standards of a European, White middle class. Researchers working from this model may view particular groups as embodiments of social problems and view social problems as puzzles that only educated outsiders are able to solve (Tuhiwai Smith, 2002).

Recently, more culturally appropriate research has challenged such notions by investigating the strengths and resilience of disempowered communities (e.g., Chataway, 2002). Educators and students can think more critically about the nature of deficit versus difference by being exposed to culturally different explanations for behavior that may not be found in standard psychology textbooks. In a nonexperimental research methods class, I introduce as one resolution to the co-
The Mixer: Introducing the Concept of Factor Analysis

Dan J. Segrist & Laura A. Pawlow

Southern Illinois University Edwardsville

Version of record first published: 05 Dec 2007

To cite this article: Dan J. Segrist & Laura A. Pawlow (2007): The Mixer: Introducing the Concept of Factor Analysis, Teaching of Psychology, 34:2, 121-123

To link to this article: http://dx.doi.org/10.1080/00986280701293222

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.tandfonline.com/page/terms-and-conditions

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
The Mixer: Introducing the Concept of Factor Analysis

Dan J. Segrist and Laura A. Pawlow
Southern Illinois University Edwardsville

This study entailed the development and implementation of a classroom activity designed to introduce students to the concept of factor analysis. We implemented the activity in both a personality theories course and a tests and measurements course. Data suggest that students learned about factor analysis from this activity, while enjoying it. Professors can easily modify this activity to meet the needs of a variety of psychology courses in which factor analysis is relevant or discussed.

Cooperative learning exercises enhance attitudes toward learning statistical concepts and facilitate interactions with peers in the classroom by moving students “away from the experience of learning as a solitary activity that is detached from the social context” (Perkins & Saris, 2001, p. 113). In demonstrating statistical concepts, Connor (2003) described activities that required students to move about the classroom. She noted that, among other benefits, such activities involved active learning and made statistical concepts more tangible and concrete.

This article describes an activity that introduces students to factor analysis by requiring them to move about the classroom and collaborate with peers. The objective of this activity is to demonstrate factor analysis as a statistical technique that allows researchers to condense and simplify a large number of variables into a few, more usable factors (Bernstein, Garbin, & Teng, 1988; Vogt, 1993).

Method

Participants

The sample consisted of 65 undergraduates enrolled in a theories of personality (n = 36) or tests and measurements class (n = 29). The majority of participants were women (85%), psychology majors (86%), and students in their junior (39%) or senior (57%) year.

Procedure

In the tests and measurements class, we used this activity in conjunction with lectures on test development and construction, whereas in the personality theories class, we introduced it prior to a discussion of trait theories of personality (e.g., Five-Factor Theory; McCrae & Costa, 1996). Prior to the activity, we developed a pool of hypothetical questionnaire items and placed them in a fishbowl. Each item related to one of the following constructs: anger and hostility (e.g., “When I am driving in rush-hour traffic, I often feel hostile”), health (e.g., “Physical exercise is an important part of my life”), anxiety (e.g., “I am easily frightened”), and optimism or pessimism (e.g., “I doubt that the future will work out well”; due to smaller class size we dropped the optimism or pessimism items from the tests and measurements course). We intentionally wrote some items to relate to more than one construct (e.g., “I often become anxious about my health”). Additionally we purposefully developed some items so that they would negatively relate to other items (e.g., “I rarely get physically violent”).

Prior to a lecture regarding factor analysis, each student selected one item from the fishbowl. Students then moved about the room, asked other students about their items, and eventually formed groups with students whom they believed had items with content related to their own. If a student believed his or her item belonged to more than one group we instructed that student that he or she must choose only one group. We cautioned students not to make decisions for group membership based on group size, personalities of group members, or the number of people they knew in the group—only on the content of their item and its relation to the items of other group members.
Once students formed groups, they reviewed members’ items and came to a consensus on an appropriate label for their group—a label they believed best reflected the nature of their items. For example, in the tests and measurements class, students ultimately formed five groups based on the item pool and came up with the labels good health, bad health, anger, anxiety, and relaxation.

After students completed this task and were still in their groups, discussion with the entire class began with these questions: (a) How did you make decisions about which group to join and which not to join? What other groups did you consider, and why did you decide not to join them? If you had an item that you believed could fit into more than one group, how did you make your decision? (b) What labels did each group come up with to represent their group? How satisfied are you that the label accurately represents your group’s items? What other labels did you consider? How did you ultimately decide on your chosen label? (c) What items in your group caused you the most difficulty in terms of incorporating them into the group label? Are there any items in your group that you think do not belong?

A lecture describing factor analysis as a data reduction technique used in test construction or the development of contemporary personality theories (depending on the course) followed this discussion. We pointed out how the in-class “factor analysis” had reduced the data from 36 items (or 29 in the tests and measurements class) to a few factors (represented by the student groups). We emphasized the idea that although students likely made their decisions about group membership based on the “face validity” of each item, factor analysis relies on statistical relations among items. Consequently, factor analysis might not group items in a way that makes intuitive or even theoretical sense. We stressed the notion, though, that factors are generally intended to reflect underlying psychological constructs.

The concept that items do not have to be similar to be related in factor analysis can be difficult to grasp. During the formation of groups, some students in our classes struggled with particular items because they negatively related to the other items in a group. During the discussion of the activity we reiterated that because factor analysis highlights statistical relations, items may have inverse relations with factors and, consequently, negative loadings. For example, in the tests and measurements class we explained that an actual factor analysis might collapse good health and bad health into a single group (e.g., health). Lecture and discussion concluded by underscoring the benefits and criticisms of factor analysis.

Results and Discussion

Students evaluated this activity on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree) in terms of the degree to which it helped them understand the purpose and mechanics of factor analysis (M = 4.34, SD = .94), how much they enjoyed the activity (M = 4.22, SD = .99), and how much they recommend that the instructor continue using the activity (M = 4.32, SD = .90). An independent-samples t test revealed no significant differences between the personality theories and tests and measurements classes for any of the evaluation items (all ps > .05).

We also compared performance between two sections of the tests and measurements class, taught by the same professor, on a multiple-choice exam item regarding the process of factor analysis. One section included the factor analysis activity, whereas the other entailed only a lecture on factor analysis. Although performance differences between the two sections on this exam item did not reach statistical significance (p = .11, Fisher’s Exact Test), differences were in the expected direction, with a 17% increase in the number of students correctly answering the exam question in the section participating in the activity.

In conclusion, this activity allows students to see and experience factor analysis “at work” while requiring interaction and collaboration with peers. We found no significant differences in activity ratings for two different courses taught by two different instructors, suggesting that a variety of professors teaching a variety of courses can effectively implement this activity. Students reported enjoying the activity and learning from it. The activity is easy to implement, involves all students, and is easy to modify. For example, an instructor can add or delete items to meet the needs of any class size. Another possible variation involves using items from an existing factor analyzed measure. At the conclusion of the exercise students can compare their in-class “factor analysis” with the actual data, an activity that should stimulate interesting discussions in terms of how and why their solution is similar to or different from the original (e.g., how specific items loaded, labels for the factors).
References


Notes

2. We thank anonymous reviewers for helpful comments on previous drafts of this article.
3. Send correspondence to Dan J. Segrist, Department of Psychology, Southern Illinois University Edwardsville, Alumni Hall, Campus Box 1121, Edwardsville, IL 62026; e-mail: dsegris@siue.edu.
“Dealing” With the Central Limit Theorem

David C. Matz & Emily L. Hause

Augsburg College,
St. Mary's College, California

Version of record first published: 17 Jul 2008

To cite this article: David C. Matz & Emily L. Hause (2008): “Dealing” With the Central Limit Theorem, Teaching of Psychology, 35:3, 198-200

To link to this article: http://dx.doi.org/10.1080/00986280802186201

PLEASE SCROLL DOWN FOR ARTICLE
We describe an easy-to-employ, hands-on demonstration using playing cards to illustrate the central limit theorem. This activity allows students to see how a collection of sample means drawn from a nonnormally distributed population will be normally distributed. Students who took part in the demonstration reported it to be helpful in understanding the theorem, enjoyable, and worthwhile. An analysis of open-ended exam questions indicated that students who participated in this demonstration possessed greater understanding of the central limit theorem than those who participated in an alternate demonstration and those who witnessed a standard presentation on the central limit theorem that included no demonstration.

The normal curve is a challenging concept for students in introductory statistics courses, as are sampling, sampling distributions, and probability (see Dyck & Gee, 1998). It is hardly surprising that the central limit theorem (CLT), which combines all of these concepts, is often difficult for students to comprehend (see Aberson, Berger, Healy, Kyle, & Romero, 2000). However, because the CLT is critical to the comprehension of probability, sampling, hypothesis testing, and the normal curve, students must develop an understanding of it.

Illustrations of the CLT are often necessarily dense. Mathematical proofs of the theorem are available (e.g., Kallenberg, 1997) but are often too intricate to be useful in clarifying the concept to introductory students. A better option is demonstration. Some of the more complex demonstrations employ probability (e.g., Zerbolio, 1989) and require students to recognize the shape of the normal curve in the infrequently occurring large deviations between sample means and the population mean and the more frequently occurring small deviations.

Other somewhat more concrete demonstrations rely on numbers generated by the lecturer that are then selected into samples (e.g., Johnson, 1986). The lecturer starts with raw or imagined data in a nonnormal distribution and then creates samples of size n to create a sampling distribution of means. Plotting these means using different ns demonstrates the shape of the sampling distribution. This technique has the advantage of working through the creation of a series of sampling distributions, which helps to emphasize the distinction between samples and sampling distributions.

A disadvantage of this type of demonstration is that students who are struggling with the concepts surrounding the CLT may also find generated numbers somewhat abstract. Instead of constructing or imagining a population, we use decks of playing cards to serve as the original distribution. Playing cards have the dual advantages of being concrete objects with a clearly nonnormal distribution of values and of cueing students to think in terms of probability.

Method

Participants

Sixty-five students enrolled in three sections of an introductory statistics course participated in this study in exchange for course credit. Students enrolled in one section of the course (n = 25) participated in our demonstration of the CLT in which we used playing cards to represent the samples drawn from a population (PC demonstration, see next section). Students in a second section (n = 20) participated in an alternate demonstration in which we used blocks of
random numbers to represent samples (RN demonstration, as described previously; Johnson, 1986). Those in the remaining section (n = 20) did not participate in a demonstration (no demonstration control), but instead witnessed a standard presentation on the CLT.

Procedure for the PC Demonstration

After discussing the properties of the CLT, the instructor placed an empty grid on an overhead. The grid was 13 units wide (and numbered at the bottom) by 10 units high. The instructor then asked students what a distribution of two decks of cards, ace through king, would look like if plotted on the grid (assuming that aces = 1, jacks = 11, queens = 12, and kings = 13). Students quickly recognized that eight instances of each denomination constituted a flat or platykurtic distribution. The instructor marked this flat distribution on the grid and then asked what a distribution of sample means drawn from this “population” of cards would look like. Students are often reluctant to accept that a distribution of sample means from such a population will approach a normal curve with a mean equal to the mean of means (in this case 6.5, or 7 if rounded to the nearest whole number). After being queried about the shape of a sampling distribution, the instructor dealt hands or samples of three cards to each student and asked them to calculate the mean, rounded to the nearest whole number. When students finished calculating the means for all of the samples, the instructor plotted those means on a second grid.

The plot of sample means begins to resemble a normal curve as more and more means are plotted. The instructor can modify the size of the initial population (i.e., the number of decks of cards used) and the sample size (i.e., the size of the “hands” dealt to each student) to accommodate classes of any size and illustrate the impact of changes in n on the shape of the sampling distribution. Twenty-five to 30 samples seems to be a minimum to achieve a plot that resembles a normal curve. A discussion of the demonstration and the CLT ensues.

Evaluation Materials

Student ratings. Students who participated in a demonstration indicated on 7-point scales ranging from 1 (strongly disagree) to 7 (strongly agree) their reactions to five questions. These questions addressed whether the demonstration in which they participated helped them better understand the CLT, would help them remember the CLT, was enjoyable, should be used again in the future, and was a beneficial use of class time.

Exam question. Approximately 2 weeks after the demonstrations, students responded to an open-ended exam question asking them to explain or define the CLT. We used the responses to this question as our measure of demonstration effectiveness. Two independent raters who were blind to condition scored the responses on a 7-point scale ranging from 1 (demonstrates little or no understanding of the CLT) to 7 (demonstrates a thorough understanding of the CLT; α = .85). Responses that received low scores typically omitted large portions of the definition (e.g., did not mention that the distribution of sample means will approach a normal curve as n increases or that the mean of the sampling distribution would be equal to the population mean). Responses that received high scores included (at minimum) reference to both the shape of the sampling distribution and the relation of the mean of the sampling distribution to the population mean. We used the average of the raters’ responses in the analyses.

Results

Student Ratings

Students reported that they enjoyed both demonstrations, found them beneficial to their understanding and remembering of the CLT, and recommended continued use of the demonstrations. Furthermore, students reported that the demonstrations were a beneficial use of class time. One-sample t tests indicated that all means were significantly different from the scale midpoint (ps < .01). Although evaluations of both demonstrations were favorable, evaluations of the PC demonstration were generally higher (more positive) than were evaluations of the RN demonstration (see Table 1).

Exam Question

We calculated mean scores on the exam question for students in each of the three course sections (PC demonstration, RN demonstration, no demonstration control). Results of an ANOVA indicated a significant effect for course section, F(2, 62) = 3.99, p < .03, η² = .11. Planned least significant difference follow-up tests indicated that students who participated in the PC demonstration (M = 4.22, SD = 1.79) showed a greater
Table 1. Mean Student Ratings of the Demonstrations

<table>
<thead>
<tr>
<th>Item</th>
<th>PC Demonstration</th>
<th>RN Demonstration</th>
<th>t(40)</th>
<th>Cohen's d</th>
</tr>
</thead>
<tbody>
<tr>
<td>This demonstration helped me understand the central limit theorem.</td>
<td>5.78</td>
<td>5.00</td>
<td>1.97</td>
<td>0.61</td>
</tr>
<tr>
<td>This demonstration will help me remember the central limit theorem.</td>
<td>5.96</td>
<td>4.89</td>
<td>3.24*</td>
<td>0.99</td>
</tr>
<tr>
<td>I enjoyed this demonstration.</td>
<td>6.26</td>
<td>4.79</td>
<td>4.18*</td>
<td>1.27</td>
</tr>
<tr>
<td>I would recommend this demonstration be used again in the future.</td>
<td>6.43</td>
<td>5.21</td>
<td>4.12*</td>
<td>1.23</td>
</tr>
<tr>
<td>I felt this demonstration was a waste of time.</td>
<td>1.43</td>
<td>1.89</td>
<td>1.53</td>
<td>0.46</td>
</tr>
</tbody>
</table>

Note. Students reported their responses on a scale ranging from 1 (strongly disagree) to 7 (strongly agree). *Indicates a significant difference between row means (p < .01). All means are significantly different from the scale midpoint (p < .01).

understanding of the CLT than those who participated in the RN demonstration (M = 2.90, SD = 1.42, p < .02, d = .82) and those who witnessed a standard presentation on the CLT (M = 3.20, SD = 1.71, p < .05, d = .58). Scores for those who participated in the RN demonstration did not differ significantly from those who witnessed a standard presentation (p > .50, d = .19).

Discussion

We showed that the use of playing cards to demonstrate the properties of the CLT can be a powerful learning tool that helps students comprehend and remember this often confusing concept. Although students perceived both the PC and RN demonstrations as worthwhile and enjoyable and believed that the demonstrations would help them understand and remember the properties of the CLT, student responses were generally more favorable toward our PC demonstration than the RN demonstration. As evidence of the effectiveness of our PC demonstration, students who participated in the demonstration showed enhanced understanding of the CLT relative to those who participated in the RN demonstration and those who witnessed a standard presentation of CLT material.

The uniqueness of this demonstration lies in part in its corporeal nature. Because most students are familiar with playing cards, they can readily visualize how a population of cards would be distributed. Such a method does not require students to imagine a hypothetical or abstract distribution—a task that is not always easy for students. Other advantages of this demonstration derive from the involvement of all students and the visual representation of the generated sampling distribution. Students do not infer the shape of a sampling distribution from calculated probabilities or merely observe an instructor generating a sampling distribution. Rather, they are able to actively calculate and plot sample means and watch the curve approach normal as n increases.

References


Notes

1. We thank Stacy Freiheit and Bridget Robinson-Riegler for their help in scoring responses.
2. Send correspondence to David C. Matz, Department of Psychology, Augsburg College, Campus Box 44, 2211 Riverside Avenue, Minneapolis, MN 55454; e-mail: matz@augsburg.edu.
A Palatable Introduction to and Demonstration of Statistical Main Effects and Interactions

Andrew N. Christopher
Albion College

Pam Marek
Kennesaw State University

Because concrete explanations in a familiar context facilitate understanding, we illustrate the concept of an interaction via a baking analogy to provide students with food for thought. The demonstration initially introduces the concepts of independent and dependent variables using a chocolate chip cookie recipe. The demonstration provides an opportunity to distinguish between the main effect of an ingredient and the interactions between ingredients on taste perception. Student performance on exam questions suggests that this demonstration effectively helps students understand the concept and interpretation of an interaction.

Statistical information is often abstract and not readily applied to situations outside the classroom (Conners, Mccown, & Roskos-Ewoldsen, 1998); thus, teaching statistics effectively to undergraduate students can be a challenge. However, solving problems that highlight statistical thinking in an everyday context (e.g., regression to the mean) enhances students’ statistical reasoning (Lawson, Schwiers, Doellman, Grady, & Kelnhofer, 2003). Additionally, Guttmanova, Shields, and Caruso (2005) emphasized the importance of making statistical concepts more concrete and more accessible to students, suggesting that teachers use definitional formulas rather than computational formulas to foster conceptual understanding. This article describes a demonstration that uses an everyday context to promote conceptual understanding of the concepts of main effects and interactions using concrete examples of variables that are familiar to students.

Conducting the Demonstration

Table 1 describes materials and preclass preparation suggestions. We introduce the concepts of independent variables (ingredients), levels (amounts of each ingredient), and dependent variable (taste perception) after we complete the demonstration when students make the connection between the concrete exemplars and abstract concepts (variables and levels). We begin the demonstration by saying that there are individual differences in food preferences. Some people, we say, enjoy eating baking soda. When we pull a box of baking soda out of a brown bag and offer it to students to eat, they refuse to taste it. We repeat the procedure for flour and vanilla extract. By this point in the demonstration, many students recognize that these ingredients are needed to make cookie dough. We ask students for the other ingredients and pull them out of the bag as students identify them. If all students do not realize what the ingredients have in common, we explain they are needed for chocolate chip cookie dough.

As we announce the decision to make cookies, we remove the seal from the baking soda box and dump the entire box into the small mixing bowl. Students react quickly and negatively, indicating that the amount (level) of baking soda will affect the taste and that, in a standard cookie dough recipe, much less baking soda is needed. We then dump the baking soda from the bowl into an empty sandwich bag and ask the class if anyone has the correct amount of baking soda with them. Of course, our student assistant does, and we
Table 1. Materials and Preclass Preparation

<table>
<thead>
<tr>
<th>Supplies</th>
<th>Ingredients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large mixing bowl</td>
<td>Baking soda (1 ts.)</td>
</tr>
<tr>
<td>Small mixing bowl</td>
<td>Flour (2.25 cups)</td>
</tr>
<tr>
<td>Mixer with beaters</td>
<td>Salt (1 ts.)</td>
</tr>
<tr>
<td>Mixing spoon</td>
<td>Granulated sugar (.75 cup)</td>
</tr>
<tr>
<td>Plastic sandwich bags (6)</td>
<td>Brown sugar (.75 cup)</td>
</tr>
<tr>
<td>Measuring spoons</td>
<td>Vanilla extract</td>
</tr>
<tr>
<td>Brown grocery bags (2)</td>
<td>Eggs (2)</td>
</tr>
<tr>
<td></td>
<td>Butter (2 sticks)</td>
</tr>
<tr>
<td></td>
<td>Chocolate chips</td>
</tr>
</tbody>
</table>

Precourse Preparation (approximately 10 min)
1. Put designated amount of baking soda, flour, salt, granulated sugar, and brown sugar in individual sandwich bags.
2. Put the six sandwich bags into one of the brown grocery bags. The empty sixth sandwich bag is used for storing excess baking soda after dumping the entire box of it into the small mixing bowl.
3. Seal packages of baking soda, flour, salt, granulated sugar, and brown sugar.
4. Put sealed packages in second brown grocery bag together with mixing bowls, mixer, and measuring spoons.
5. Ask one student to assist in demonstration, providing ingredients in sandwich bags as needed.

Put that amount of it in a small mixing bowl. Next, students respond to inquiries about the appropriate levels of flour and salt in the recipe, which we then add to the small mixing bowl. Using a large mixing bowl, we continue the procedure, asking about the levels of granulated sugar, brown sugar, vanilla, and eggs, getting the bagged ingredients from the student assistant, and adding them to the bowl together with 1 teaspoon of vanilla and two eggs. At this point, we ask if anyone would like to eat the mixture (but no one does). After mixing the ingredients in the large bowl for about 30 sec, we ask about the appropriate amount of butter to add, and after adding it, we again ask if anyone would like to eat the mixture (again no one does).

However, after we combine the ingredients in the small bowl with those in the large bowl, some students are eager to taste the dough, particularly if we have added a bag of chocolate chips to the resulting mixture. This shift in students’ willingness to sample the mixture provides a segue into a discussion of independent and dependent variables, main effects (one ingredient’s effect on taste perception), and interactions (different levels of different ingredients’ effects on taste perception). Students can readily articulate how combining different levels of the various ingredients changes taste perceptions. However, it is imperative to emphasize that simply changing the level (percentage) of one ingredient (e.g., baking soda) in a standard-size cookie dough recipe does not illustrate an interaction, but rather, a main effect of one independent variable on taste perceptions. The large amount of baking soda with which the demonstration began was only inappropriate because of the relatively smaller amounts of the other ingredients. In other words, the effect of the amount of baking soda on taste perception depends on the amounts of the other ingredients, illustrating an interaction. Specifically, a high level of baking soda in the standard-size cookie dough recipe results in poor taste perception, but a high level of baking soda in a relatively larger size recipe would result in good taste perception. Thus, the level of baking soda interacts with the levels of the other ingredients to affect taste perception. To illustrate this interaction, we graphed our hypothetical interaction in crossover form, with the amount of baking soda on the x-axis, quality of taste on the y-axis, and one line representing each level of batch size. Discussion has extended to how taste perception would change if we added or subtracted one particular ingredient (or omitted it entirely). The demonstration and discussion of key concepts typically requires about 25 min of class time.

For the remainder of the class session, we present and discuss results of an actual experiment using a 2 × 2 design, with results revealing a statistically significant interaction, but no significant main effects. We expand this example, usually in the next class meeting, to discuss the different possible outcomes of a 2 × 2 design (i.e., only one significant main effect, with or without a significant interaction, and two significant main effects, with or without a significant interaction), concluding with a review of the result from the actual experiment, a significant interaction but neither main effect is significant. After this conceptual overview of main effects and interactions, we then teach students likely expect to receive real chocolate chip cookies after the demonstration is complete.

1Because cookie dough contains raw eggs, we do not allow students to eat the final product in the demonstration. Additionally, teachers should be aware that students will put that amount of it in a small mixing bowl.

2We thank an anonymous reviewer for this pedagogical improvement to the demonstration.
how to calculate a $2 \times 2$ ANOVA by hand, run this analysis on SPSS, and interpret the resulting output.

**Evaluation**

To evaluate the extent to which the cookie demonstration enhanced students’ ability to explain main effects and interactions more generally, we compared response scores on three open-ended questions. We administered these questions to two different classes, one exposed to the demonstration (Fall 2006) and one not exposed to the demonstration (Spring 2007) on two separate occasions, once at the beginning of the second semester of a two-semester course sequence on research design and again about 2 weeks after conducting the cookie demonstration. The same teacher taught both classes. Other than the cookie demonstration, students in both classes viewed the same presentation of interaction outcomes and completed similar hand calculations, homework assignments, and SPSS lab assignments related to interactions.

We graded the three questions on a scale ranging from 1 to 7, with higher scores indicating greater mastery. Although we as teachers graded these questions for the purpose of assigning exam grades, we had two senior undergraduate psychology majors code these particular responses to help assure objectivity for the purpose of assessing the effectiveness of the demonstration.3

The first question asked students to explain the difference between a “main effect” and an “interaction” in an experiment. The second question asked how an interaction could be statistically significant if the two main effects were nonsignificant. The final question required students to interpret a line graph showing an interaction without using “statistical jargon.” We formed composite variables by averaging together the three items on the pretest ($\alpha = .73$) and the three items on the posttest ($\alpha = .72$). We ran a 2 (demonstration group: received demonstration or did not receive demonstration) $\times$ 2 (time of administration: predemonstration and postdemonstration) ANOVA, with repeated measures on the second factor. As revealed in Figure 1, there was a significant interaction, $F(1, 49) = 4.26$, $p = .044$, $\eta^2 = .08$. Although the groups did not differ on the pretest measure, $t(49) = .23$, $p = .81$, students exposed to the demonstration performed better on the posttest measure than did students not exposed to the demonstration, $t(49) = 2.50$, $p = .016$.

**Caveats and Conclusions**

Our data suggest that the demonstration enhanced students’ understanding of interactions. In evaluating its effectiveness, we held potentially confounding variables, such as instructor and other interaction-related materials, constant. Yet, as is the case with any design involving intact groups, we acknowledge that other individual differences might have influenced results. Additionally, as is true for most demonstrations that are assessed via exam questions, assessment was targeted at evaluating understanding of the general underlying concept rather than the demonstration itself.

Insofar as preparation for the demonstration, instructors might be concerned with the amount of time involved in shopping for ingredients and preclass preparation. If time commitments are problematic, we suggest using a “simpler” recipe (e.g., boxed brownie mix). Furthermore, if it is not feasible to actually bring butter and eggs to class, instructors might use yellow and white modeling clay replicas. The important points are that students are able to identify the effect of the amount of a specific ingredient in a single recipe (main effect) on taste perception and understand how the effect of the level of a specific ingredient on taste perception depends on the levels of the other ingredients (interaction). Student exam performance hinted that time involved in this demonstration could be time well invested. In sum, using a widely known cookie dough recipe to introduce the notions of statistical main

3The results of our assessment did not change when we analyzed grades assigned by the instructor. In fact, instructor grades correlated .92 with those of our undergraduate coders.
effects and interactions appears to be an enjoyable and pedagogically effective strategy.

References


Notes

1. We presented portions of this research at the 17th Annual Southeastern Conference on the Teaching of Psychology, Atlanta, GA, and at the 28th National Institute on the Teaching of Psychology, St. Petersburg Beach, FL.

2. We thank Jennifer Christianson and Jason Streeter for their help coding the assessment data. We thank Mark I. Walter and Jared W. Keeley for their insights on conducting this demonstration. Randolph Smith and three anonymous reviewers provided valuable suggestions during the editorial process.

3. Send correspondence to Andrew N. Christopher, Department of Psychology, Albion College, 611 E. Porter Street, Albion, MI 49224; e-mail: achristopher@albion.edu.
A Demonstration of the Analysis of Variance Using Physical Movement and Space
William J. Owen and Paul D. Siakaluk
Teaching of Psychology 2011 38: 151
DOI: 10.1177/0098628311411779

The online version of this article can be found at:
http://top.sagepub.com/content/38/3/151

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Society for the Teaching of Psychology

Additional services and information for Teaching of Psychology can be found at:

Email Alerts: http://top.sagepub.com/cgi/alerts
Subscriptions: http://top.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Jun 14, 2011
What is This?
A Demonstration of the Analysis of Variance Using Physical Movement and Space

William J. Owen and Paul D. Siakaluk

Abstract
Classroom demonstrations help students better understand challenging concepts. This article introduces an activity that demonstrates the basic concepts involved in analysis of variance (ANOVA). Students who physically participated in the activity had a better understanding of ANOVA concepts (i.e., higher scores on an exam question answered 2 months after the classroom demonstration) than did students who simply observed the activity.

Keywords
classroom demonstration, ANOVA, statistics, active learning

Teaching statistics to undergraduate psychology students presents unique challenges. Conners, Mccown, and Roskos-Ewoldsen (1998) highlighted four such challenges: (a) motivating students to learn the subject content, (b) reducing math anxiety, (c) the existence of performance extremes (i.e., a large difference between high and low achievers), and (d) emphasizing that the learning of statistics is important for future psychology classes and, we would argue, for everyday experience. With respect to motivating students and reducing performance extremes, Conners et al. suggested that instructors use active learning strategies (i.e., instructional methods used for promoting engaged cognitive activity as opposed to mere passive reception) and concrete problems to enhance the mastery of statistical concepts. Following this suggestion, Segrist and Pawlow (2007) utilized active learning and concrete examples to enhance students’ understanding of factor analysis. In this article, we introduce a classroom activity that utilizes physical movement and space to elucidate basic concepts associated with analysis of variance (ANOVA). This classroom activity was purposely designed to provide a conceptual summary underlying the logic of ANOVA rather than a computational summary of the mathematics involved in ANOVA. We chose this method of explaining the basic concepts of ANOVA because the textbook (Aron, Aron, & Coups, 2006) that we used in our undergraduate statistics course focused on conceptual formulae.

Psychologists have shown that learning activities utilizing physical movement in space serve to increase students’ interest and understanding of challenging concepts. For example, Connor (2003) suggested that activities requiring students to move to a spot on a number line drawn on the floor to indicate their response for a particular question (e.g., do you like statistics?) enhanced student interest for concepts such as the shapes of frequency distributions, variation, and central tendency. Students also found such demonstrations to be useful in terms of their statistical understanding. Connor also had students move to a spot on a quadrant to demonstrate the concepts of correlation, regression, and scatterplots.

The ANOVA statistical technique is commonly introduced in undergraduate psychology statistics courses. Conceptually, this procedure involves partitioning experimental variance into two main sources, between-group variance and within-group variance, which are attributable to treatment effects plus between-group error and random sampling error, respectively. By utilizing an appropriate dependent variable (e.g., height), variability can be measured and represented in an embodied manner. Because bodies vary in height and it is likely that individuals use their own bodies as a reference point for comparing variation in height, we capitalized on this embodied knowledge by using height as the dependent variable. The basic conceptual principles of ANOVA were therefore illustrated via physical movement and space in the demonstration described below.

Demonstration
Levels of a Single Factor
This demonstration was used in a 3rd-year statistics course in psychology, which is an entry-level statistics course in our department. In all, 40 students were enrolled in the course. When introducing ANOVA in the classroom, we had students...
select one of three responses to the question, “What is your favorite psychology course?” Students were asked to choose their favorite psychology course out of a list comprising social, clinical, and cognitive psychology. Based on their responses, six students were then randomly selected to represent each of the three levels of this factor (i.e., favorite course). Within each group, we asked students to arrange themselves according to height. At this point, we discussed the concepts of a single factor with multiple levels.

Factors Versus Levels
We had ensured an equal mix of men and women in each of the three groups. By including a mix of men and women in each group, this demonstration can be expanded to illustrate a two-factor ANOVA. In the mixed-gender groupings, students could see that there was one factor (i.e., favorite course) with three levels. Asking the women in each group to step forward allowed us to show that we could have two factors (i.e., favorite course and gender), one factor with three levels and the other factor with two levels. Following this explanation, the men and women went back to their original groupings for a one-factor ANOVA.

The rest of this demonstration focused on a one-way ANOVA. We addressed the question of whether one’s favorite course, the independent variable, had an impact on height, the dependent variable. The 18 students’ heights, three groups of six individuals, represented the Xis in the formulae. We then demonstrated the concept of the F ratio. Imagining a different experimental scenario or more precise measurements, we asked the students who represented the group means to take one step closer to their groups. As such, the distance between the sample and the group mean was reduced and the distance between the group means and the GM indicating any differences between a group mean and the GM by placing his hands on the students’ heads to visually illustrate the differences. This source of variability was attributable to treatment effects plus between-group error. For this demonstration treatment effects were the result of favorite class whereby different classes cause different types of growth (in class, students were told to imagine a hypotheti- cal situation in which cognitive classes cause cognitive growth and an increase in height because of growing brains, social classes cause social growth and in increase in one’s girth because of growing bodies of social networks, etc.).

Between-group error effects were the result of random differences such as biological factors, socioeconomic factors, and so on. The three students representing the group means then figured the between-group variability, with the GM then standing at a distance representative of this calculation. The total variability (i.e., \(X(X_{GM})^2\)) was demonstrated by having the group means sit down on the floor so the audience could focus on the total distance between individuals and the GM. We then visually illustrated how the individuals’ heights (representing the raw scores) differed from the GM’s height. We also verbally highlighted that the total variability equals the within-group variability plus the between-group variability.

F Ratio
We then demonstrated the concept of the F ratio. Although we did not calculate the F ratio for this demonstration, we discussed how the within-group variability and between-group variability are used to figure mean square within and mean square between. In addition, we discussed how changes to the within-group variability and/or between-group variability would affect the F ratio. Imagining a different experimental scenario or more precise measurements, we asked the students who represented the group means to take one step closer to their groups. As such, the distance between the sample and the group mean was reduced and the distance between the group means and the GM was increased. We discussed that the point of reducing the distance between the group means and the individual scores represented a reduction in random sampling error, whereas the point of increasing the distance between the group means and the GM represented an increase in the magnitude of the treatment effects. We then discussed ways in which researchers may reduce the within-group variability and/or increase the between-group variability. It is important that we emphasized to the students the importance of these possible ways of changing experimental variability (e.g., more precise
measurments, validity of the dependent measures) that serve to increase the likelihood that one would find a significant $F$ ratio (i.e., a significant treatment effect).

**Evaluation and Discussion**

One week after the classroom demonstration, students evaluated the ANOVA demonstration by responding to a short questionnaire that asked about the perceived usefulness of the activity in terms of enhancing understanding and student interest in learning about ANOVA. Students responded to each question using a 5-point scale from 1 (strongly disagree) to 5 (strongly agree). They also indicated whether they were a participant (i.e., part of the demonstration) or an observer. A total of 16 participants and 14 observers completed the survey. In addition, students were asked a question on a midterm exam (given 2 months after the classroom demonstration) where they were asked to graphically illustrate the concepts from a one-way ANOVA. It is important that the individual who conducted all the data analyses and who graded the exam question was blind to which students were in which group. One-sample $t$ tests were used to determine if responses differed from a neutral endorsement ($M = 3$). Overall, student responses indicated that the activity increased their perceived understanding ($M = 4.03, SD = 0.85$), $t(29) = 6.65, SEM = 0.16, p < .01$, $\eta^2 = .60$, enhanced their ability to visualize key ANOVA concepts ($M = 4.10, SD = 0.76$), $t(29) = 7.94, SEM = 0.14, p < .01$, $\eta^2 = .69$, and was more interesting than learning the concepts from the textbook ($M = 4.33, SD = 0.76$), $t(29) = 9.63, SEM = 0.14, p < .01$, $\eta^2 = .76$.

Independent samples $t$ tests were used to compare differences between those who participated and those who observed the demonstration. There was a difference between those who participated ($M = 4.38, SD = 0.50$) and those who simply observed ($M = 3.64, SD = 1.01$) in terms of the perceived usefulness of the classroom activity in understanding the concepts associated with ANOVA, $t(28) = 2.57, SEM = 0.29, p < .05$, $\eta^2 = .19$. This finding was concordant with results on a midterm question worth 5 points asking students to graphically illustrate the components into which a score’s deviation from the GM can be divided. Participants scored higher ($M = 4.50, SD = 0.63$) on the ANOVA question than did observers ($M = 3.57, SD = 1.08$), $t(28) = 2.90, SEM = 0.32, p < .01$, $\eta^2 = .24$.

This demonstration captured the basic conceptual principles of ANOVA by relying on physical movement and space. Of the 40 students in the class, 22 participated in the demonstration. Those students who were not selected to participate were considered the observers. The core concepts illustrated in the activity included factors, levels, group means, GM, within-group variability, between-group variability, and the $F$ ratio.

We used height as a dependent variable in this study because it is easy for people to visualize variations in height; however, one could use other measures, such as foot size, hand size, or arm length. The important consideration for alternative dependent variables is that they allow for a visually apparent measurement (e.g., using one’s hands to measure variability in arm length).

In conclusion, this classroom activity provides an interesting and concrete way for undergraduate psychology students to learn about ANOVA. Our students found it interesting and enjoyable, and it enhanced their understanding of subtle concepts, especially of the partitioning of experimental variance. As such, we recommend that instructors provide opportunities for their students to directly participate in such active learning demonstrations. It is important that this activity can be modified to (a) demonstrate the concepts of factorial ANOVA, planned comparisons, and post hoc tests; (b) incorporate follow-up calculation examples to illustrate within-subject and between-subject variability (Johnson, 1989; Scuitto, 2000); and (c) contrast this method with other methods of active learning (see Mayer, 2004). As noted, teaching statistics to undergraduate psychology students poses unique challenges. However, we think that these challenges allow opportunities for instructors to design and create innovative approaches to teaching.

**Acknowledgments**

We want to thank Drs. Penny Pexman (University of Calgary) and Elizabete Rocha (UNBC) as well as two anonymous reviewers for providing very helpful comments on earlier versions of this article. We would also like to thank the teaching assistant, Tammy Klassen, for her assistance with this project.

**Declaration of Conflicting Interests**

The authors declared no potential conflicts of interests with respect to the authorship and/or publication of this article.

**Funding**

The authors received no financial support for the research and/or authorship of this article.

**Notes**

1. Knowledge gained through bodily experience is commonly referred to as embodied cognition (see Barsalou, 1999) or grounded cognition (see Barsalou, 2008).
2. It is important for instructors of statistics to note that the effects of (a) reduced within-group variability and (b) increased between-group variability are, in fact, independent mechanisms that can serve to increase the size of the $F$ ratio.
3. The midterm question related to this activity was, “Using these points of reference—scores, cell means, grand mean—graphically illustrate all the components that a score’s deviation from the grand mean can be divided into when conducting a one-way ANOVA. Label each component that you illustrate (5 marks).” The teaching assistant was given a copy of a figure from the textbook that showed all the components of a one-way ANOVA to use as a marking rubric.
4. We would like to thank an anonymous reviewer for suggesting this recommendation.
References


Bios

**William J. Owen** is a cognitive psychologist whose interests are in teaching, faculty development, and research of basic visual word recognition. I have enjoyed teaching statistics and cognitive psychology for both undergraduate and graduate students.

**Paul D. Siakaluk** is a cognitive psychologist whose primary research interests are in studying language and conceptual processing from an embodied cognitive perspective. I have taught a variety of psychology courses at both the undergraduate and graduate levels, including introductory, cognitive, evolutionary, and statistics.
t for Two: Using Mnemonics to Teach Statistics

Daniel R. Stalder¹ and Elizabeth A. Olson¹

Abstract
This article provides a list of statistical mnemonics for instructor use. This article also reports on the potential for such mnemonics to help students learn, enjoy, and become less apprehensive about statistics. Undergraduates from two sections of a psychology statistics course rated 8 of 11 mnemonics as significantly memorable and helpful in learning statistics. Undergraduates rated the 3 remaining mnemonics as helpful after excluding students who did not recall those mnemonics (beyond scale midpoint). Other measures indicated a relatively high regard for the overall use of statistical mnemonics. In particular, mnemonics were rated as motivating and fun. Students also reported moderate belief that mnemonics could reduce statistics anxiety. The variety of positive findings suggests strong promise for statistical mnemonics.

Keywords
mnemonics, statistics anxiety, statistics education

Statistics is one of the most anxiety-producing subjects for college students, including social science majors (e.g., Kher, Molstad, & Donahue, 1999; Onwuegbuzie & Wilson, 2003). In fact, interest in psychology negatively correlates with interest in math (Stalder, 2002). Finding ways to make statistics material more accessible seems especially important for psychology majors who need statistics to succeed in the research methods course, to conduct independent research, and to understand research articles.

Because statistics and research methods are often core courses in the undergraduate social science curriculum (Perlman & McCann, 2005), improving students’ relationship with statistics is a critical first step in improving student engagement across the curriculum. Among helpful learning methods, mnemonics have received support for statistics, research methods, and other material (e.g., Carney & Levin, 2008; Higbee, 1996; Lakin, Giesler, Morris, & Vosmik, 2007; Saber & Johnson, 2008; Stalder, 2005; VanVoorhis, 2002). Unlike other effective strategies in statistics and research methods which require considerable instructor effort (Dunn, Smith, & Beins, 2007), most mnemonics are easily added to existing material. The research-based benefit of mnemonics weighed against this low cost strongly suggests trying mnemonics. To this end, we provide a list of statistical mnemonics, including first-letter mnemonics (e.g., t tests compare two groups; Stalder, 2005) and those involving the keyword method (Carney & Levin, 2008). We also report student ratings on the degree to which such mnemonics might help students learn, enjoy, and feel less anxiety over material.

Method
Sixty-one undergraduates (17 men, 44 women; age M = 22.2, SD = 4.8) completed anonymous semester-end surveys for extra credit from two sections of an introductory psychology statistics course at a Midwestern university. The surveys reviewed 11 mnemonics provided during the semester (see Table 1) and asked four questions about each (1 = not at all, 7 = a great degree): “To what degree” (a) “was this mnemonic helpful in learning this information,” (b) “did this mnemonic make learning this information easier or faster,” (c) “did this mnemonic make the information easier to recall during homework or tests,” and (d) “do you recall this mnemonic from the semester?” We averaged the first three items to create one perceived helpfulness measure for each mnemonic (zs ranged from .87 to .98). Then, participants answered 11 questions about their overall perceptions of mnemonics (1 = not at all, 7 = a great degree) (see Table 2). Nine of these questions assessed the perceived value of mnemonics, including the degree to which mnemonics increased learning, motivation, and fun, should be used by other instructors, and helped to reduce statistics anxiety. Two other items assessed use of mnemonics by other instructors.

¹ University of Wisconsin–Whitewater, Whitewater, Wisconsin, USA

Corresponding Author:
Daniel R. Stalder, Department of Psychology, University of Wisconsin–Whitewater, 800 W. Main Street, Whitewater, WI 53190-1790.
Email: stalderd@uww.edu
For each measure, we compared mean student scores against scale midpoints using one-sample $t$ tests. Participants significantly recalled 8 of the 11 mnemonics and rated those same mnemonics as significantly helpful using the three-item helpfulness measure (see Table 1). Considering participants who recalled the 3 remaining mnemonics above the scale midpoint (approximately half the sample), perceived helpfulness ratings significantly (or marginally) exceeded the scale midpoint ($p_s < .01; p = .079$ for the ANOVA mnemonic).

Overall perceptions of mnemonics indicated relatively high regard (see Table 2). In particular, participants generally reported that mnemonics improved learning and motivation (including motivation to create their own mnemonics), made the material more fun, and should be provided more by other instructors. Ratings of mnemonic use by other instructors were relatively low. The mean rating of 4.0 for how much the mnemonics reduced statistics anxiety did not exceed the scale midpoint. However, given the prevalence and degree of statistics anxiety (e.g., Onwuegbuzie & Wilson, 2003), a comparison

Table 1. Mean and Median Scores (and Standard Deviations) for Recall and Perceived Helpfulness of Mnemonics

<table>
<thead>
<tr>
<th>Mnemonics</th>
<th>Degree Recalled</th>
<th>Perceived Helpfulness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M$</td>
<td>$Mdn$</td>
</tr>
<tr>
<td>1. The &quot;t&quot; in &quot;t test&quot; stands for &quot;two,&quot; and t tests compare two groups (including one sample and one population).</td>
<td>6.4***</td>
<td>7.0</td>
</tr>
<tr>
<td>2. The &quot;F&quot; in &quot;F test&quot; stands for a &quot;few,&quot; &quot;four,&quot; or &quot;five,&quot; and F tests compare three or more groups.</td>
<td>6.5***</td>
<td>7.0</td>
</tr>
<tr>
<td>3. The &quot;r&quot; in r tests stands for &quot;relationship,&quot; and r tests (or correlations) measure the relationship between variables.</td>
<td>5.7***</td>
<td>6.0</td>
</tr>
<tr>
<td>4. &quot;P&quot; stands for &quot;parameters&quot; and &quot;population,&quot; and parameters refer to the population; similarly, &quot;s&quot; stands for &quot;statistics&quot; and &quot;sample,&quot; and statistics refer to a sample.</td>
<td>4.5*</td>
<td>5.0</td>
</tr>
<tr>
<td>5. To remember the four types of scales, think of stereotypical statisticians wearing glasses and shirts with wrinkles, and there may be a humorous sign on the wall in the lab saying, &quot;NO IRoning allowed&quot; (N = __, O = __, I = __, R = __).</td>
<td>4.7**</td>
<td>5.0</td>
</tr>
<tr>
<td>6. &quot;O&quot; stands for &quot;ordinal&quot; and &quot;Olympics&quot;; the ordinal scale is used in the Olympics.</td>
<td>5.6***</td>
<td>6.0</td>
</tr>
<tr>
<td>7. To help with Step 6 in hypothesis testing (i.e., that we want $t_{comp}$ and $F_{comp}$ to be as big as possible, to reject $H_0$), think of a huge computer screen, like those big-screen TVs at Best Buy...</td>
<td>5.1***</td>
<td>5.0</td>
</tr>
<tr>
<td>8. To remember that the F test in Ch.11 is also known as an ANOVA (analysis of variance), think of a supernova (an exploding star) and how there are many, many stars, like the many, many groups an ANOVA can statistically compare.</td>
<td>4.2</td>
<td>4.0</td>
</tr>
<tr>
<td>9. To remember that $f_{0}$ (e.g., 2) comes before $f_{w}$ (e.g., 18) in parentheses when writing out the F report, such as $F(2,18) = 6.62, p &lt; .05$, use alphabetical order: b comes before w in the alphabet.</td>
<td>5.4***</td>
<td>6.0</td>
</tr>
<tr>
<td>10. To remember the formula for the &quot;mean,&quot; think about a mean group/sum of foXes pouncing on a Newt [animal illustration also provided]: $\sum_{i=1}^{n} X_i \overline{X}$</td>
<td>4.3</td>
<td>4.0</td>
</tr>
<tr>
<td>11. To remember which column to use, column B or C, in the z table (to find an area between 0 and your z score or an area beyond your z score), notice that B stands for between and C stands for corner (of the graph).</td>
<td>4.2</td>
<td>5.0</td>
</tr>
</tbody>
</table>

Note. Means and medians can range from 1 to 7. Higher scores represent greater degree recalled or perceived helpfulness. Means that were significantly higher than the midpoint of the scale (4.0) using a one-sample t test are indicated by the $p$ value superscripts; the effect size column pertains to perceived helpfulness ts.

* $p < .05$. ** $p < .01$. *** $p < .001$. 

Results

For each measure, we compared mean student scores against scale midpoints using one-sample $t$ tests. Participants significantly recalled 8 of the 11 mnemonics and rated those same mnemonics as significantly helpful using the three-item helpfulness measure (see Table 1). Considering participants who recalled the 3 remaining mnemonics above the scale midpoint (approximately half the sample), perceived helpfulness ratings significantly (or marginally) exceeded the scale midpoint ($ps < .01; p = .079$ for the ANOVA mnemonic).
Table 2. Mean and Median Scores and Standard Deviations for Overall Perceptions of Mnemonics

<table>
<thead>
<tr>
<th>Perception Measure</th>
<th>M</th>
<th>Mdn</th>
<th>SD</th>
<th>η²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceived helpfulness in...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>learning and retaining material</td>
<td>5.8***</td>
<td>6.0</td>
<td>1.0</td>
<td>.79</td>
</tr>
<tr>
<td>improving course performance</td>
<td>5.4***</td>
<td>6.0</td>
<td>1.1</td>
<td>.62</td>
</tr>
<tr>
<td>getting motivated to start studying</td>
<td>4.6**</td>
<td>5.0</td>
<td>1.5</td>
<td>.12</td>
</tr>
<tr>
<td>making the material more fun to learn</td>
<td>4.7***</td>
<td>5.0</td>
<td>1.6</td>
<td>.18</td>
</tr>
<tr>
<td>Actual use of the mnemonics during the semester</td>
<td>5.4***</td>
<td>5.0</td>
<td>1.2</td>
<td>.59</td>
</tr>
<tr>
<td>Recalled use by other college instructors</td>
<td>2.9***</td>
<td>2.0</td>
<td>1.6</td>
<td>.34</td>
</tr>
<tr>
<td>Recalled use by other math or statistics instructors</td>
<td>2.6***</td>
<td>2.0</td>
<td>1.6</td>
<td>.44</td>
</tr>
<tr>
<td>View on whether...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>other math or statistics instructors should provide mnemonics</td>
<td>6.0***</td>
<td>6.0</td>
<td>1.2</td>
<td>.73</td>
</tr>
<tr>
<td>other students should use mnemonics</td>
<td>6.1***</td>
<td>6.0</td>
<td>1.3</td>
<td>.72</td>
</tr>
<tr>
<td>participants themselves would now create their own mnemonics</td>
<td>5.6***</td>
<td>6.0</td>
<td>1.5</td>
<td>.55</td>
</tr>
<tr>
<td>the mnemonics helped to reduce statistics anxiety</td>
<td>4.0</td>
<td>4.0</td>
<td>1.6</td>
<td>.00</td>
</tr>
</tbody>
</table>

Note. Means and medians can range from 1 to 7. Higher scores represent greater degree. Means that were significantly different than the midpoint of the scale (4.0) using a one-sample t test are indicated by the p value superscripts; the effect size column pertains to these t scores.

Discussion

Based on multiple measures, students reported significant regard for most of the specific mnemonics and for mnemonics in general in learning statistics. In particular, students thought that mnemonics increased learning, motivation, and fun and should be provided more by other instructors. Having fun seems particularly important to address statistics anxiety (e.g., Lomax & Moosavi, 2002). Students also reported moderate belief that the provided mnemonics reduced anxiety. Such beliefs strengthened among students who better recalled the mnemonics, possibly lending some validity to such beliefs. Given the prevalence and degree of statistics anxiety (e.g., Onwuegbuzie & Wilson, 2003), a tool that receives even a mid-scale overall rating for reducing anxiety seems worth trying. Students also reported that other instructors, including math and statistics instructors, use mnemonics infrequently. These results might suggest that mnemonics are a relatively untapped resource in current post-secondary education (see also Stalder, 2005). Carney and Levin (2008) suggested that part of this low instructor use might be due to instructors’ misconceptions about mnemonics, which Carney and Levin hoped their research could address.

As a limitation, the design was nonexperimental and did not demonstrate actual improvements in performance due to mnemonics. Other research has provided such evidence for statistics, research methods, and other material (e.g., Carney & Levin, 2008; Higbee, 1996; Lakin et al., 2007; Saber & Johnson, 2008; Stalder, 2005; VanVoorhis, 2002), but we did not collect such data in the current work. Feelings of motivation and fun probably need to be self-reported in any case and might play a vital role in reducing statistics anxiety, which would improve performance (Onwuegbuzie & Wilson, 2003). Among benefits of the current design, the real-life classroom context breeds greater validity and generalizability of results, particularly because of the statistics anxiety that might spike during an actual statistics class. Moreover, the effect sizes for many of the students’ positive impressions were quite large (see Tables 1 and 2), in some cases over three times the conventional “large” effect size (Vacha-Haase & Thompson, 2004). Taking into account that statistics students are usually not very upbeat about the course, sometimes delaying their enrollment (Kher et al., 1999; Onwuegbuzie & Wilson, 2003) with potential negative consequences (Freng, Webber, Blatter, Wing, & Scott, in press), such large effects become even more meaningful. As recommended by Vacha-Haase and Thompson (2004) in evaluating the meaning of effect sizes, we compared present effect sizes with those of similar self-report measures in previous studies (e.g., Lakin et al., 2007; Stalder, 2005) and found present effect sizes to be much larger.

In sum, the variety and size of positive findings suggest strong potential for mnemonics to improve statistics learning and provide even clearer evidence for improving the overall (subjective) statistics course experience. Nearly all of the 11 mnemonics (Numbers 1-10 in Table 1) also cover material necessary to know for the research methods course that usually directly follows introductory statistics in the social science major. Whatever the cause of the apparent low instructor use of mnemonics, we believe that most of the mnemonics provided in this article can be easily integrated into statistics and research methods courses and will prove useful to instructors.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.
Funding
The authors received no financial support for the research, authorship, and/or publication of this article.

Acknowledgments
We thank Isabel Davis and Kyle Anderson for their valuable assistance.

References
**Student Reactions**

To test the effectiveness of the bobber demonstration, I set up the bobber model in my graduate multivariate statistics class and lectured on multivariate distributions. One week later, I asked students to take a surprise quiz and to complete a short questionnaire.

The quiz included four questions. Thirteen of 14 (93%) students correctly answered this question: How would the bobbers be configured if three variables were all negatively correlated? Twelve of 14 (86%) students correctly answered this question: Using data bobber terms, describe Mahalanobis distance. Eleven of 14 (79%) students correctly answered the next two questions: Using data bobber terms, describe the position of the regression line, and why does the shape of the bobbers look like a football?

The questionnaire included two items: The bobber model helped me understand multiple regression, and the bobber model helped me understand three-dimensional data sets much better than a two-dimensional drawing would have. Students responded to each item on a scale ranging from 1 (strongly disagree) to 7 (strongly agree). Students agreed that the bobbers helped them understand multiple regression (M = 5.7, SD = 1.33) and they thought that the three-dimensional model was more helpful than a two-dimensional line drawing (M = 6.8, SD = .43).

Perhaps the greatest benefit of the bobber model is the low fear, walk-through, nonmathematical approach to regression. In my experience, once students understand regression visually, they can more easily move on to calculations and formulas. The formulas make more sense once students see what the formulas explain.

---

**Reference**


---

**Note**

Send correspondence to Daniel A. Sachau, Department of Psychology, Minnesota State University, Mankato, MN 56001; e-mail: sachau@mankato.msus.edu.

---

**Teaching Meta-Analysis in the Undergraduate Experimental Psychology Course**

David K. Dodd

*Washington University, St. Louis*

In this article, I provide a model for teaching meta-analysis in the undergraduate experimental psychology course. I recommend certain prerequisites for both the instructor and students, outline a lecture, and make suggestions for a student exercise. Evaluative data show that the presentation, along with the in-class exercise that gives students practice in using meta-analytic procedures, is effective in increasing students’ knowledge of meta-analysis.

Meta-analysis represents a relatively new topic for undergraduate courses in experimental psychology and research methods. (Hereafter, I use experimental psychology synonymously with research methods.) Although analytic procedures for combining research findings from many studies date back at least to 1931 (see Rosenthal, 1984), it was not until the classic study of Smith and Glass (1977) that meta-analysis clearly emerged as a new approach that challenged the traditional literature review. Nonetheless, my informal review of several experimental psychology texts reveals that most texts cover meta-analysis only superficially, if at all. Students enter my experimental psychology course with little, if any, knowledge of meta-analysis. The purpose of this article is to offer instructors a model for presenting meta-analysis to students in a lecture–laboratory format that can be covered in approximately 90 min.

**Prerequisites**

Instructors of undergraduate experimental psychology courses may be reluctant to teach meta-analysis for two reasons. First, they may not feel adequately trained to teach meta-analysis. Because of the relatively recent emergence of meta-analysis, instructors trained before 1985 probably received no formal exposure to the topic, and current graduate programs may not consistently teach meta-analysis. Second, instructors may believe that meta-analysis is too complicated for undergraduate students and that graduate programs...
should be responsible for teaching it. My argument against this position is that advanced undergraduate psychology majors who conduct literature searches will almost inevitably encounter meta-analyses, and it is important that they understand enough about the topic so they can complete their literature reviews with confidence.

The instructor who feels unprepared to teach meta-analysis can learn the basics in a relatively brief time. Reading a chapter on meta-analysis at the introductory level will give a basic overview. I highly recommend the chapter in Judd, Smith, and Kidder (1991), written at a level for beginners, and the one in Rosenthal and Rosnow (1991), written at a slightly higher level. As a follow-up to this reading, I recommend a more thorough coverage of the topic, such as Rosenthal (1984) or Wolf (1986), both of which are comprehensive without being intimidating to nonmathematicians. Finally, the instructor should read at least one recent meta-analysis, paying special attention to the statistical analysis.

Certain prerequisites are also necessary for students. First, students must know the basics of inferential statistics, especially $p$, $r$, $F$, and $t$. They must also have a basic understanding of the concept of effect size. Finally, it would be desirable for the instructor to review the advantages of standard scores, such as $z$. Students in my course have previously taken statistics, but I devote a substantial part of one lab session to effect size early in the semester and do not present meta-analysis until late in the semester, after further review of statistics has occurred.

Lecture

I present a 45-min lecture at the beginning of class. The main components are definition of meta-analysis, comparison with traditional (narrative) literature reviews, identification of appropriate studies, and more advanced concepts such as the “apples versus oranges” issue, blocking, and the “file drawer” problem. I present one or two examples of meta-analytic findings that I think would be particularly interesting or applicable to students. One example I have found to be effective is a meta-analysis conducted by Mullen and Felleman (1990), who studied the relation of “tripling” in college dormitories to student adjustment. (Tripling is the assigning of three residents to a room designed for two, usually because of overcrowding.)

Another interesting example concerns gender differences in social smiling. By blocking (separating) studies according to age of participants, Hall (1984) was able to conclude that there is no gender difference in smiling among infants or children, but that there is a moderately strong difference among adults, with women smiling more than men. Using a nearly identical set of studies, Hall and Halberstadt (1986) reported a follow-up meta-analysis with the same conclusion. These studies provide the basis for an effective exercise that I now use with my class.

In-Class Exercise

To give students practice in basic meta-analytic procedures, I conduct an in-class exercise that requires about 30 to 40 min. Prior to the session, I assemble approximately 15 research articles (see Appendix) to address a single hypothesis. Sources for these articles can be a meta-analytic review (e.g., Hall, 1984) or the instructor’s own files. I provide students with simple formulas, learned during an earlier class session, for converting common statistics to effect size ($r$) and copies of a table for transforming Pearson’s $r$ to Fisher’s $z$ (commonly available in most upper level statistics books).

Students work in small groups of two or three to encourage collaborative learning. Each group receives two to four articles. I instruct students to read the abstract of each article carefully, to read the method for an operational definition of the key variable and for sample size and other characteristics, and to locate the critical statistic ($t$, $F$, $\chi^2$, or $r$) and the accompanying $p$ in the results. From the formulas and table provided, students calculate the effect size ($r$) for each article and convert it to Fisher $z$. On the board in the front of the room, students categorize the studies according to age of sample (adult vs. children) and record the year of publication, sample size, $r$, Fisher $z$, and $p$. Students then make brief (30 to 60 sec) presentations of their articles, including notable characteristics of the sample, unusual operational definitions, and any difficulties in interpreting the findings. At that point, I average the Fisher $z$ values, convert this average back to $r$, and lead a discussion about whether the results of our in-class exercise support the research hypothesis.

Besides demonstrating the procedures of meta-analysis, this exercise provides opportunities for further class discussion. Students find that obtaining an effect size can be difficult because results are sometimes reported only in terms of $p$, especially if a test is only secondary to the study’s main focus. One article, published more than 65 years ago (Ding & Jersild, 1932), drew a conclusion about smiling without reporting any inferential statistics, instead citing only behavior rates, percentages, and averages. I point out that both the availability of statistics and expectations regarding the reporting of results have changed dramatically since then. Finally, I try to make the important point that different researchers studying a single hypothesis use varying operational definitions and samples. Although this point
illustrates the apples and oranges problem, a common criticism of meta-analysis, I emphasize that including diverse studies in a meta-analysis actually is valuable because it permits an evaluation of the generality (i.e., external validity) of the findings.

Evaluation

I have made the meta-analysis presentation to six sections of my experimental psychology class (91 students). To evaluate the presentation, I administered a four-item questionnaire both at the beginning and end of the class period. Students responded to all items on a 7-point scale ranging from 1 (strongly disagree) to 7 (strongly agree). The following pre- and postpresentation means were obtained:

- I have a basic understanding of what meta-analysis is; pre $M = 2.8$ ($SD = 1.7$), post $M = 6.3$ ($SD = 0.9$).
- I understand the basic statistics used to conduct a meta-analysis; pre $M = 2.0$ ($SD = 1.2$), post $M = 6.1$ ($SD = 1.0$).
- I know what is meant by the "file drawer" problem; pre $M = 1.5$ ($SD = 1.2$), post $M = 6.3$ ($SD = 1.0$).
- I feel I know enough about meta-analysis to assist a knowledgeable researcher in conducting an analysis; pre $M = 1.5$ ($SD = 1.0$), post $M = 5.2$ ($SD = 1.3$).

All of the differences between pre- and postpresentation means were statistically significant ($p < .001$). I caution readers that these data reflect student impressions rather than objective measures of knowledge.

I administered two additional items at the end of the class period to evaluate the overall effectiveness of the presentation. In response to "this presentation and exercise have been effective in teaching me the basics of meta-analysis," 91% of the class endorsed the item at a level of 5 or better on the 7-point scale ($M = 6.0$, $SD = 1.1$). Similarly, 87% of the class endorsed the item "I recommend that this presentation/exercise be included in future courses in experimental psychology" ($M = 5.9$, $SD = 1.1$). Written comments by several students suggest that they remained confused by the statistical procedures of meta-analysis, such as the Fisher $z$-transformation, the Stouffer combined $Z$, and the fail-safe $N$. It may be unrealistic to expect students to master the several statistical methods of meta-analysis in one class period. During my most recent presentation, I simplified the statistics by focusing on effect size and only briefly mentioning statistics related to the comparison of the findings.

Conclusions

This presentation on meta-analysis has proven to be a valuable addition to my experimental psychology course. Students increase their knowledge of meta-analysis and gain experience in meta-analytic techniques during the in-class exercise. I hope this exposure to meta-analysis will serve students well in advanced undergraduate studies and in graduate programs.

References


References


Appendix

Research Articles Used in the Classroom Demonstration of Meta-Analysis


in learning the concepts from 1 (not at all useful), to 7 (extremely useful) and whether they would recommend using the demonstration in the future (yes or no). The mean rating of usefulness for the sample (N = 41) was 5.79 (SD = .93), with 66% of the students giving a rating of 6 or 7. When asked whether I should use the demonstration again, 40 of the 41 students responded “yes.” In sum, I have found this demonstration to be a useful tool for illustrating the complex interrelation of effect size, random error, power, and errors in hypothesis testing.

Reference


Note

Send correspondence to Mark J. Sciutto, Department of Psychology, Westminster College, New Wilmington, PA 16172; e-mail: sciuttmj@westminster.edu.

Floating Data and the Problem With Illustrating Multiple Regression

Daniel A. Sachau
Minnesota State University, Mankato

In this article, I describe a technique for introducing students to multiple regression using an easily constructed, large-scale, 3-dimensional model. The model helps instructors explain multivariate normal distributions, centroids, Mahalanobis distance, multivariate outliers, and regression.

It is relatively easy to illustrate simple regression at a chalkboard. Draw a horizontal X-axis predictor, a vertical Y-axis criterion, a few data points, a regression line, and there it is. Illustrating regression becomes considerably more difficult, however, when you add an additional predictor. Instructors who discuss multiple regression often ask students to imagine another axis emerging out from the chalkboard. The instructors then introduce the idea of data points with three, rather than two, coordinates and ask students to imagine the data points “floating” out in the classroom. Once data points leave the chalkboard, students often begin to struggle with demonstrations of regression.

In an effort to introduce the basic concepts of multiple regression, I create a large-scale, three-dimensional regression model using the classroom walls and floor. That is, instead of using imaginary axes, I create concrete axes by making the length of the right wall correspond to an X-axis predictor, the length of the adjoining wall correspond to a Z-axis predictor, and the height of the walls correspond to the Y-axis criterion. Then I bring in data points.

I use fishing bobbers to represent data points. After removing the furniture from half the classroom, I use fishing line and thumbtacks to suspend 50 red and white bobbers from the ceiling so that the pattern of bobbers represents three positively correlated and normally distributed variables. In other words, I arrange the bobbers in the shape of a large football. The football is tilted so that one of the points of the football is near the floor in the corner of the room (the junction of the X-, Y-, and Z-axes). The other end of the football is near the ceiling at the center of the rectangular room.

Next, I print large letters (X, Y, Z) and three sets of numbers (1 to 7) and tape them to the walls. If the classroom is constructed from concrete blocks, the joints of the blocks provide convenient points for locating axes coordinates. It takes about 30 min to set up the demonstration.

The bobber model is similar to a model described by Kirby (1976). Kirby’s model involved pegboard, rods, and spheres. Although less portable than Kirby’s model, the bobber model allows students to walk inside the distribution.

I use an employee selection example to start the discussion of multiple regression. I ask students to choose a job, a performance criterion, and two predictors. For simplicity, I ask students to select, or make up, variables with a 7-point range. For example, students might try to predict supervisory ratings of a clothing store manager’s performance using years of retail experience and assessment center performance as the predictors. Students usually come up with much more interesting or exotic examples.

Teaching Points

Centroid

The centroid is the point defined by the means of the three variables. It floats in the center of the football. For illustration, almost any bobber that is different from the data point bobbers will work nicely as a centroid.

Multivariate Normality

I ask students why three normally distributed and positively correlated variables take the shape of a football. Students have to think about how the normally distributed variables, each of which is thicker in the middle than it is on the ends, combine to form the football shape.

Mahalanobis Distance

Mahalanobis distance is the distance between a data point and the centroid (a bobber and the unique bobber). Like the centroid, Mahalanobis distance is tricky to explain but easy to point to. Mahalanobis distance is useful for defining an outlier.

Multivariate Outlier

Outliers can be a problem because outliers have a disproportionately high degree of leverage on the position of the regression line. Outliers, thus, may need to be eliminated. A
multivariate outlier is a data point that is not particularly unusual when you consider any one variable but is unusual when you consider two or more variables. To illustrate a multivariate outlier, I suspended a bobber well within the range of the X, Y, and Z variables but well outside the distribution defined by the combination of the X, Y, and Z variables. The bobber represents an employee who scores low on X and Z predictors but high on the performance rating (Y). The outlier bobber hangs near the corner of the room and is clearly outside the football. I can easily point out how the data bobber is not a distant outlier for the X predictor, the Z predictor, or the Y criterion, but is an outlier for the XY, ZY, and XYZ combinations.

Regression

I introduce the concept of least-squares regression by pointing out a line that runs through the centroid and minimizes the distance between the line and the data points. Here it is handy to have some form of pole or rod to position among the bobbers.

I allow students to use the model to predict scores on the criterion. For any given set of predictor scores, students can generally make a fairly accurate prediction simply by walking into the distribution and guessing the criterion value.

Student Reactions

To test the effectiveness of the bobber demonstration, I set up the bobber model in my graduate multivariate statistics class and lectured on multivariate distributions. One week later, I asked students to take a surprise quiz and to complete a short questionnaire.

The quiz included four questions. Thirteen of 14 (93%) students correctly answered this question: How would the bobbers be configured if three variables were all negatively correlated? Twelve of 14 (86%) students correctly answered this question: Using data bobber terms, describe Mahalanobis distance. Eleven of 14 (79%) students correctly answered the next two questions: Using data bobber terms, describe the position of the regression line, and why does the shape of the bobbers look like a football?

The questionnaire included two items: The bobber model helped me understand multiple regression, and the bobber model helped me understand three-dimensional data sets much better than a two-dimensional drawing would have. Students responded to each item on a scale ranging from strongly disagree to strongly agree. Students agreed that the bobber model helps them understand multiple regression (M = 5.7, SD = .43) and they thought that the three-dimensional model was more helpful than a two-dimensional line drawing (M = 6.8, SD = .43).

Perhaps the greatest benefit of the bobber model is the low fear, walk-through, nonmathematical approach to regression. In my experience, once students understand regression visually, they can more easily move on to calculations and formulas. The formulas make more sense once students see what the formulas explain.

Reference


Note

Send correspondence to Daniel A. Sachau, Department of Psychology, Minnesota State University, Mankato, MN 56001; e-mail: sachau@mankato.msus.edu.

Teaching Meta-Analysis in the Undergraduate Experimental Psychology Course

David K. Dodd
Washington University, St. Louis

In this article, I provide a model for teaching meta-analysis in the undergraduate experimental psychology course. I recommend certain prerequisites for both the instructor and students, outline a lecture, and make suggestions for a student exercise. Evaluative data show that the presentation, along with the in-class exercise that gives students practice in using meta-analytic procedures, is effective in increasing students’ knowledge of meta-analysis.

Meta-analysis represents a relatively new topic for undergraduate courses in experimental psychology and research methods. (Hereafter, I use experimental psychology synonymously with research methods.) Although analytic procedures for combining research findings from many studies date back at least to 1931 (see Rosenthal, 1984), it was not until the classic study of Smith and Glass (1977) that meta-analysis clearly emerged as a new approach that challenged the traditional literature review. Nonetheless, my informal review of several experimental psychology texts reveals that most texts cover meta-analysis only superficially, if at all. Students enter my experimental psychology course with little, if any, knowledge of meta-analysis. The purpose of this article is to offer instructors a model for presenting meta-analysis to students in a lecture–laboratory format that can be covered in approximately 90 min.

Prerequisites

Instructors of undergraduate experimental psychology courses may be reluctant to teach meta-analysis for two reasons. First, they may not feel adequately trained to teach meta-analysis. Because of the relatively recent emergence of meta-analysis, instructors trained before 1985 probably received no formal exposure to the topic, and current graduate programs may not consistently teach meta-analysis. Second, instructors may believe that meta-analysis is too complicated for undergraduate students and that graduate programs...
Senior Year Retention of Methods and Statistics Concepts
James Scepansky and David M. Carkenord
Teaching of Psychology 2004 31: 9
DOI: 10.1207/s15328023top3101_3

The online version of this article can be found at:
http://top.sagepub.com/content/31/1/9
This article describes an empirical investigation of whether performance on a senior-level assessment of methodological and statistical knowledge was related to elapsed time since students took prerequisite methods and statistics courses. In a senior-capstone course, 50 students completed a 50-item multiple-choice assessment measuring their knowledge of methods and statistics. We also measured elapsed time (number of months) since students completed Research Methods and Quantitative Methods (statistics). Results revealed that performance on the assessment showed no significant correlation with elapsed time from the Research Methods course or with elapsed time from the Quantitative Methods course. On a self-report survey, however, students re-
ported they believed the elapsed time negatively affected their performance.

Most undergraduate psychology programs require majors to take statistics and research methods courses (Friedrich, Buday, & Kerr, 2000; Perlman & McCann, 1999). Messer, Griggs, and Jackson (1999) reported that programs in their sample required a mean of 1.9 statistics and methods courses, with the most prevalent pattern being the requirement of one statistics and one methods course. Given this widespread trend, it is worth examining student acquisition and retention of material from these courses. Research testing students’ memory for material from methods courses has indicated that students appear to internalize methods content more than content in other courses. Specifically, students in methods courses were more likely to report that they “just knew” answers to test questions, as opposed to “remembering” the specific answers (Conway, Gardiner, Perfect, Anderson, & Cohen, 1997; Herbert & Burt, 2001). Conway et al. suggested that this schematization might be due to the focused nature of the content of a methods course. If such schematization does occur, then students should retain a reasonable amount of their knowledge for an extended period after the course is completed. Herbert and Burt assessed student memory 6 weeks after the conclusion of a methods course and found no significant deterioration of memory. Our study expanded on this work by examining the effects of much larger spans of elapsed time on memory for content knowledge of both methods and statistical concepts.

Origin of This Study

Based on recommendations made during a recent program self-assessment at our institution, we now formally assess senior-psychology majors for retention of material from our Research Methods and Quantitative Methods (statistics) courses. Prior to taking the assessment, many students complained that they took the two courses long ago and expressed fear that the delay would hinder their assessment performance. We were curious as to whether performance on the assessment was in fact related to the amount of elapsed time since students took the courses.

Method

Participants

Participants were 50 senior-psychology majors enrolled in two sections of Senior Seminar (Fall, 2001: n = 21; Spring, 2002: n = 29) who had completed the Research Methods and Quantitative Methods course earlier in their course progression. Approximately 56% (n = 28) of our sample completed Quantitative Methods before Research Methods, 32% (n = 16) completed Research Methods first, and 12% (n = 6) completed the courses concurrently.

Materials

The assessment was a 50-item multiple-choice test measuring knowledge of methods and statistical concepts. The assessment contained some factual questions, but the majority of items were applied in nature (e.g., read an experimental scenario and identify the independent and dependent variable and appropriate statistic).

Procedure

During the first class meeting we told students about the assessment, that they would complete it the following week (we teach Senior Seminar as a once-a-week class), and that it would constitute 25% of their course grade. During the third class meeting students received their assessment scores and then completed a survey measuring their perceptions of the assessment process. We asked students how they believed the time interval between completion of the two courses and the current assessment affected their assessment performance. For each course, students provided a rating on a 7-point scale ranging from 1 (time interval very definitely hurt my performance) to 7 (time interval very definitely helped my performance). Finally, students estimated the number of hours they had reviewed for the assessment in the preceding week.

Results

An independent t test comparing the assessment scores (items correct) of students in Fall 2001 (M = 34.48, SD = 4.78) and Spring 2002 (M = 35.66, SD = 4.78) revealed no significant difference, t(48) = 0.86, p > .05. Thus we collapsed the data for analysis. To obtain a measure of elapsed time since students had taken the courses, the first author (who taught all sections of the courses) examined his records to determine the number of months since students had completed Quantitative Methods (M = 16.18, SD = 8.46, range = 1 to 33) and Research Methods (M = 14.12, SD = 9.05, range = 1 to 32). We believed that the large standard deviations and ranges indicated adequate variation on the elapsed time measure, so we correlated the assessment scores with the elapsed time. Results revealed no significant correlation between the scores and elapsed time from the Quantitative Methods course, r(48) = .14, p > .05, nor elapsed time from the Research Methods course, r(48) = -.13, p > .05. However, analysis of the survey items assessing students’ perceptions of the effects of the elapsed time indicated that students believed the time interval negatively affected their performance (Research Methods, M = 2.65, SD = 1.69; Quantitative Methods, M = 2.33, SD = 1.33). The number of hours that students reviewed in the preceding week (M = 2.57, SD = 1.76) was not correlated with performance on the assessment, r(48) = -.01, p > .05. Finally, scores on the assessment were positively correlated with grade earned, on a 4-point scale (A = 4, B = 3, etc.), in Research Methods, r(48) = .60, p < .05, and Quantitative Methods, r(48) = .42, p < .05.

We conducted further analyses to determine whether the number of psychology classes taken after completing Quantitative Methods and Research Methods was related to scores on the assessment. We identified the number of courses taken after completion of the two methods courses (but before the semester during which Senior Seminar was taken) and the total number of psychology classes taken prior to the semester in which students enrolled in Senior Seminar.
Scores on the assessment were not related to the number of classes taken since completing Research Methods, \(r(48) = .02, p > .05\), the number of classes taken since completing Quantitative Methods, \(r(48) = .17, p > .05\), nor the total number of Psychology classes completed before enrolling in Senior Seminar, \(r(48) = .24, p > .05\).

Finally, we examined whether the order in which students completed the two courses was related to assessment performance. We compared assessment scores across three categories of students: those who took Research Methods first \((n = 16)\), those who took Quantitative Methods first \((n = 28)\), and those who took the two courses concurrently \((n = 6)\). A one-way ANOVA indicated a significant difference, \(F(2, 47) = 3.83, p < .05\). Specifically, students who completed Quantitative Methods first performed better on the assessment \((M = 36.68, SD = 4.49)\) than students who completed Research Methods first \((M = 33.63, SD = 3.38)\) and students who completed the courses concurrently \((M = 32.17, SD = 6.97)\).

**Discussion**

Our data indicated that the level of seniors’ methods and statistical knowledge did not correlate with elapsed time since students completed the courses, although students believed that the elapsed time had a negative impact on their performance. Our data are consistent with the propositions of Conway et al. (1997) as well as Herbert and Burt (2001), who suggested that schematization of methods content occurs at the time the course is taken. Our data suggested the same may be true of statistical content as well. What appears to matter most is whether students grasp the concepts at the time they completed the courses, as evidenced by the fact that students who earned higher grades in each class tended to earn higher scores on the senior assessment. That students believed the elapsed time since completing the courses hindered their performance is not surprising, however. Research has shown that people are not necessarily aware, at least explicitly, of the factors that influence their actions, but instead make implicit assumptions as to likely influences (e.g., Nisbett & Wilson, 1977). In this regard, it is a logical assumption that information learned long ago may not be as accessible as material learned recently.

Interestingly, we found a nonsignificant relation between the amount of time students spent reviewing for the assessment \((M = 2.57 \text{ hr})\) and assessment performance. This finding also supports the notion that content is schematized at the time of initial learning. Methodologically it would be desirable to compare the assessment scores of a class that was given advanced notice of the assessment to a class that was not given such notice to more decisively test whether student review impacts assessment performance. However, as the assessment is a required course component accounting for 25% of the course grade, we believe it would be inappropriate to “surprise” students with the assessment. Another methodological option would be to give an unannounced assessment, but not use the assessment scores as part of the course grade. Under these conditions, however, the level of students’ motivation to perform well on the assessment would be suspect.

Although not central to our investigation, we also found empirical evidence supporting an advantage for a specific sequencing of the research methods and quantitative methods courses within a curriculum. Students who took the Quantitative Methods course before the Research Methods course performed better on the assessment. One possible explanation for this finding may be that students who first learn the basic mechanics of hypothesis testing, probability, and statistical operations are better able to then schematize that information with the information they subsequently learn in the methods course. Students who complete the methods course first may not experience the same degree of consolidation of knowledge and thus not retain the knowledge long term. Another possible explanation may be that stronger students take the statistics class earlier in their course progression, and weaker students may delay taking the class due to fear and apprehension. The stronger students then, not surprisingly, also do better on the senior assessment. However, it is important to note that the majority of our students complete these two courses, regardless of order, in two successive semesters. We therefore feel confident in contending that amount of time since taking the classes, in and of itself, does not influence performance on the senior assessment.

Some limitations of our study should be noted. First, our assessment instrument was not empirically validated by previous research. However, we took the majority of the items on the assessment instrument from test banks that accompany numerous research methods, statistics, and experimental psychology texts. Also, several members of our department viewed and endorsed the assessment, suggesting that it was a face valid measure of methods and statistics concepts. A second limitation centers on the timing of the attitude survey administered to students. Because we provided students with the assessment results prior to their being asked to complete the survey, the knowledge of their performance on the assessment might have induced some bias in their ratings on the survey. Those students who did better on the assessment may not have perceived the elapsed time to be quite as detrimental as those students who did poorly on the assessment. Yet we believe that any bias that was introduced was small, in that the mean ratings on the attitude items fell between the somewhat hurt my performance and slightly hurt my performance anchor points on the scale. Even though students were aware of their performance, they still generally believed that the elapsed time hurt their performance. A potentially interesting replication of our study would have students complete the survey immediately after the assessment, before they receive their results.

**References**


---

**Notes**

1. This article was based on a poster presented at the Eastern Conference on the Teaching of Psychology, October 2001.

2. We thank three anonymous reviewers and Randolph Smith for their comments and feedback on earlier versions of this manuscript.

3. Send correspondence to James Scepansky, Department of Psychology, Cedar Crest College, Allentown, PA 18104; e-mail: jascepan@cedarcrest.edu.
Teaching of Psychology
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/htop20

Should College Algebra be a Prerequisite for Taking Psychology Statistics?
Amy E. Sibulkin & J. S. Butler

a Tennessee State University,
b Martin School of Public Policy and Administration, University of Kentucky,

Version of record first published: 17 Jul 2008

To cite this article: Amy E. Sibulkin & J. S. Butler (2008): Should College Algebra be a Prerequisite for Taking Psychology Statistics?, Teaching of Psychology, 35:3, 214-217

To link to this article: http://dx.doi.org/10.1080/00986280802286399

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.tandfonline.com/page/terms-and-conditions

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
Should College Algebra be a Prerequisite for Taking Psychology Statistics?

Amy E. Sibulkin
Tennessee State University

J. S. Butler
Martin School of Public Policy and Administration, University of Kentucky

In order to consider whether a course in college algebra should be a prerequisite for taking psychology statistics, we recorded students’ grades in elementary psychology statistics and in college algebra at a 4-year university. Students who earned credit in algebra prior to enrolling in statistics for the first time had a significantly higher mean statistics grade than those without algebra. However, prior credit in algebra did not significantly add to the prediction of statistics grades, after we controlled for GPA, sex, and year in college.

Instructors of undergraduate elementary statistics in our Psychology Department sometimes express dissatisfaction with the poor performance of many students in their sections. After observing that many students did not pass college algebra, the instructors often urge the department to change its policy from strongly recommending to requiring that students earn credit in college algebra before taking statistics. Enforcing a prerequisite is costly to faculty in terms of time and stress in dealing with dissatisfied students. The cost to students is a delay in their progress in completing their major. Therefore, we undertook a systematic analysis of the relation between completion of algebra and subsequent grade in statistics before deciding whether to make the change of creating an “enforced prerequisite,” meaning a course students must successfully complete prior to statistics.

At the national level, Perlman and McCann (1999) reviewed course requirements for a psychology major for a sample of 500 colleges and universities representing the four Carnegie categories. Of the departments that offered statistics, 68% listed Introductory Psychology as a prerequisite. The authors did not explicitly state whether they considered courses outside the department, such as algebra.

We did not find any studies on the relation between college algebra and psychology statistics grades. The most closely related work was on the relation between intermediate algebra and introductory statistics in a mathematics department (Sworder, 1997) and between intermediate algebra and introductory chemistry (Wilson, 1994). We chose Butler, Finegan, and Siegfried’s (1998) research on teaching economics as a model for our study, particularly due to their inclusion of GPA as a control variable. They predicted grades in intermediate microeconomics and macroeconomics from whether students had two prior calculus courses, as opposed to one. They controlled for GPA and several other variables related to math performance. The second semester of calculus was associated with higher grades in intermediate microeconomics but not in intermediate macroeconomics. The authors explained the results in terms of the lack of use of calculus in the teaching of intermediate macroeconomics.

As psychology statistics does require studying equations, it seems logical to assume that knowledge at the level of college algebra is important for success. However, our anecdotal evidence revealed several students who earned credit in statistics without algebra. Therefore, the purpose of this project was to estimate the independent association between grade in college algebra and grade in statistics. Given our practical interest in advising and curriculum review, as opposed to a conceptual model of performance, we included only three control variables in the model predicting statistics grade. We controlled for GPA, the variable that represents general success in college classes. We also
included class year to control for the increase in cognitive skills and motivation that more advanced students may bring to a difficult course. Finally, we included sex. Women usually have higher GPAs (e.g., Lammers, Onwuegbuzie, & Slate, 2001; Nguyen, Allen, & Fraccastoro, 2005; Strahan, 2003), possibly due to specific skills, such as coping strategies (Clifton, Perry, Stubbs, & Roberts, 2004).

Method

Participants

We identified all students from any major who enrolled in elementary psychology statistics in Fall 2001 or Spring 2002 at Tennessee State University, a public historically Black university (HBCU). Fall 2001 enrollment statistics showed that 89% of psychology majors were African American and 10% were Caucasian. Eighty percent of psychology majors were women. There were five sections of statistics each semester. We included all majors but excluded the few students who had taken a math course higher than college algebra (such as calculus), leaving a sample of 208 students.

Procedure

Given that many students took college algebra and statistics more than once due to failure or withdrawal, we examined the entire transcript of these 208 students in order to record the first semester in which the student enrolled in statistics and the grade in statistics. Regarding the question of the influence of algebra on statistics, the relevant characteristic is the grade in algebra before taking statistics, regardless of how many times the student took algebra. Therefore, we recorded the most recent grade (or withdrawal) in algebra that was prior to first enrollment in statistics. If the student’s first enrollment in algebra was during the same semester as statistics or after statistics, then the student did not have a grade in algebra prior to statistics. They were coded as zero on a dummy variable representing credit or no credit in statistics prior to algebra.

Results

For the 181 students who had a grade in algebra prior to taking statistics for the first time, the correlation between algebra grade and statistics grade was .22, and the correlations of cumulative GPA with algebra grade and statistics grade were .38 and .51, respectively. (All r's have 179 degrees of freedom and p < .001.) Given the significant correlations between GPA and both algebra and statistics grades, Table 1 shows the average grade in statistics for each grade level in algebra (A = 4, B = 3, C = 2, D = 1, F = 0) for the total sample. We also show grades separately for high-GPA and low-GPA students, in order to observe the pattern of algebra and statistics grades relative to GPA. We used the median GPA as the cutoff for high-GPA and low-GPA students (2.70 or higher was “high,” and 2.69 or lower was “low”). We included in Table 1 the group of 27 students who did not have algebra before statistics and showed their average statistics grade as well. (Of these 27 students, 11 took algebra simultaneously with statistics.)

<table>
<thead>
<tr>
<th>Grade in algebra, if taken before statistics</th>
<th>High-GPA Students</th>
<th>Low-GPA Students</th>
<th>All Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>19 3.16 1.53</td>
<td>5 1.40 0.49</td>
<td>24 2.79 1.55</td>
</tr>
<tr>
<td>B</td>
<td>33 2.94 1.30</td>
<td>22 1.41 1.37</td>
<td>55 2.33 1.53</td>
</tr>
<tr>
<td>C</td>
<td>31 2.81 1.15</td>
<td>34 1.59 1.31</td>
<td>65 2.17 1.38</td>
</tr>
<tr>
<td>D</td>
<td>4 2.25 1.79</td>
<td>20 1.55 1.07</td>
<td>24 1.67 1.25</td>
</tr>
<tr>
<td>F</td>
<td>4 2.25 1.30</td>
<td>8 1.25 1.20</td>
<td>12 1.58 1.32</td>
</tr>
<tr>
<td>W</td>
<td>1 2.00 0.00</td>
<td>1 2.00 0.00</td>
<td>1 2.00 0.00</td>
</tr>
<tr>
<td>No algebra before statistics</td>
<td>12 2.33 1.37</td>
<td>15 1.20 1.11</td>
<td>27 1.70 1.36</td>
</tr>
<tr>
<td>A through D</td>
<td>87 2.91 1.34</td>
<td>81 1.52 1.24</td>
<td>168 2.24 1.47</td>
</tr>
<tr>
<td>F or no algebra</td>
<td>16 2.31 1.36</td>
<td>24 1.25 1.13</td>
<td>40 1.68 1.33</td>
</tr>
<tr>
<td>Total</td>
<td>103 2.82 1.36</td>
<td>105 1.46 1.22</td>
<td>208 2.13 1.46</td>
</tr>
</tbody>
</table>

Note. The university does not award plus and minus grades.
We also combined the first 4 rows of the table into the group “A through D” and combined the “F,” “W,” and “No algebra” rows into the group “F or no algebra” in order to compare the mean statistics grade of students who earned credit in algebra with those who did not. For the overall sample, the “A through D” algebra group earned significantly higher statistics grades than the “F or no algebra” group (A through D grade $M = 2.24$, F or no grade $M = 1.68$), $t(206) = 2.36$, $p < .05$, $\eta^2 = .023$, with an effect size of .39 using Cohen’s $d$ based on the overall standard deviation, 1.46.

Within each algebra grade level, including not having algebra at all, the mean statistics grade was higher for the high-GPA group than the low-GPA group, suggesting an independent association of GPA and statistics grade. Therefore, we regressed statistics grade on the following variables in Step 1: (a) cumulative GPA, (b) sex ($1 =$ woman), and (c) whether student was a junior or senior ($1 =$ Yes). The reference group of “Sophomore and other” contained 83 students. For Step 2 we entered (d) algebra grade ($0 =$ F, W, or no algebra; $1 =$ A, B, C, or D, if taken before statistics). The unstandardized regression coefficients appear in Table 2.

The variables entered in Step 1 produced an $R^2$ of 28%, $F(3, 204) = 30.32$, $p < .001$. The $R^2$ increased by a nonsignificant 1% after entering algebra grade in Step 2, indicating that algebra grade did not have a significant association with grade in statistics, after controlling for the other variables. GPA had a strong positive association with statistics grade. The coefficient for GPA of 1.26 indicates that a 1-unit change in GPA (e.g., from 2.7 to 3.7 on a 4-point scale) was associated with an increase in statistics grade of over one letter grade. Being a woman independently added to the prediction of statistics grade; women were estimated to score about two-thirds of a letter grade higher than men.

### Table 2. Summary of Hierarchical Regression Analysis Predicting Grade in Statistics (N = 208)

<table>
<thead>
<tr>
<th>Variable</th>
<th>$B$</th>
<th>$SEB$</th>
<th>$t$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Step 1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cumulative GPA</td>
<td>1.30</td>
<td>0.20</td>
<td>6.50*</td>
</tr>
<tr>
<td>Woman</td>
<td>0.59</td>
<td>0.24</td>
<td>2.52*</td>
</tr>
<tr>
<td>Junior or senior</td>
<td>0.01</td>
<td>0.17</td>
<td>0.03</td>
</tr>
<tr>
<td><strong>Step 2</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cumulative GPA</td>
<td>1.26</td>
<td>0.20</td>
<td>6.38**</td>
</tr>
<tr>
<td>Woman</td>
<td>0.60</td>
<td>0.24</td>
<td>2.53*</td>
</tr>
<tr>
<td>Junior or senior</td>
<td>−0.04</td>
<td>0.18</td>
<td>−0.22</td>
</tr>
<tr>
<td>A – D in algebra, if taken before statistics</td>
<td>0.33</td>
<td>0.23</td>
<td>1.45</td>
</tr>
</tbody>
</table>

*Note. Regressions are estimated using STATA with heteroscedasticity-consistent standard errors; $R^2 = .28$ for Step 1; $\Delta R^2 = .01$ (ns); $F(4, 203) = 24.82$, $p < .001$. $^*p < .05$. $^**p < .001$."

Discussion

Many students begin college requiring remediation in math (Parsad, Lewis, & Greene, 2003), and many students fail or withdraw from college algebra (e.g., Hagerty, 2005; Small, 2006; Thiel, n.d.; Twigg, 2004). Therefore, many students are on track to enroll in statistics before they have earned credit in college algebra. Should advisors or instructors allow them to enroll in statistics? Clearly, we could not randomly assign students to groups that did or did not pass algebra before enrolling in statistics for the first time. Therefore, we can not directly observe the causal effect of grade in algebra on grade in statistics. However, most policy decisions use nonexperimental data.

It is true that absent of control variables, average statistics grades were significantly higher for students who earned an A through D grade in algebra before enrolling in statistics for the first time. The policy implication regarding an algebra prerequisite rests on how one perceives the relevance of the control variables, particularly GPA, the key policy-relevant variable. The regressions are based on the premise that GPA is a general indicator of the ability and motivation that students apply to a course. GPA was a significant predictor of statistics grade. A further premise of the regression is that a significant coefficient for algebra would reflect knowledge particular to completing an algebra course that adds to performance in statistics, beyond the general performance indicated by GPA. The regression showed that completing algebra did not significantly add to the prediction of statistics grade. If one thinks that usefulness of algebra as a prerequisite requires that it “stand alone” as a predictor of statistics, then one would not recommend a prerequisite of algebra. Alternatively, if one thinks it is invalid to separate the specific performance in algebra from performance indicated by the more general GPA, then one would recommend a prerequisite of algebra, based on the bivariate association of algebra with statistics grades.

Of course, there are many other variables related to grade in statistics, such as amount of time...
spent studying, quality of the instructor, level of test anxiety, and competing extracurricular activities. We emphasize again that the present study is policy driven, designed to address the value of an enforced prerequisite, and is not a behavioral model to explain statistics grade.

Finally, we consider the nature of elementary statistics, as typically taught. Usually, students do not solve equations for an unknown. However, some instructors do teach the course with theory and algebra, and our results do not generalize to these settings.

References


Note

Send correspondence to Amy E. Sibulkin, Department of Psychology, Tennessee State University, 3500 John Merritt Boulevard, Nashville, TN 37209; e-mail: asibulkin@tnstate.edu.
The Role of Statistics and Research Methods in the Academic Success of Psychology Majors: Do Performance and Enrollment Timing Matter?
Scott Freng, David Webber, Jamin Blatter, Ashley Wing and Walter D. Scott

*Teaching of Psychology* 2011 38: 83
DOI: 10.1177/0098628311401591

The online version of this article can be found at:
http://top.sagepub.com/content/38/2/83

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Society for the Teaching of Psychology

Additional services and information for *Teaching of Psychology* can be found at:

Email Alerts: http://top.sagepub.com/cgi/alerts
Subscriptions: http://top.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Mar 23, 2011

What is This?
The Role of Statistics and Research Methods in the Academic Success of Psychology Majors: Do Performance and Enrollment Timing Matter?

Scott Freng,1 David Webber,2 Jamin Blatter,3 Ashley Wing,1 and Walter D. Scott1

Abstract

Comprehension of statistics and research methods is crucial to understanding psychology as a science (APA, 2007). However, psychology majors sometimes approach methodology courses with derision or anxiety (Onwuegbuzie & Wilson, 2003; Rajecki, Appleby, Williams, Johnson, & Jeschke, 2005); consequently, students may postpone enrollment (Onwuegbuzie, 2004). We examined the importance of methodology course performance (i.e., grade) and timing of enrollment in these courses for the academic success of psychology majors. After controlling for number of courses taken, relevant American College Test (ACT) scores, and prior Grade Point Average (GPA), we found that methodology course grade and timing of enrollment predicted upper-division psychology GPA. In addition, methodology course grade predicted performance on an assessment of knowledge in psychology. We discuss implications of these results for teaching and advising.

Keywords

research methods, statistics, student performance

If asked what course they most dread taking, many undergraduate psychology students will say statistics or research methods. Trepidation regarding statistics and research methods may come from statistics anxiety (Onwuegbuzie & Wilson, 2003) or students’ view that methodology courses (e.g., statistics and research methods) are less interesting than topical courses, such as abnormal psychology (Rajecki, Appleby, Williams, Johnson, & Jeschke, 2005). A potential consequence of this trepidation is that students may delay enrollment in statistics and research methods courses until late in their academic careers (Onwuegbuzie, 2004). This postponement is unfortunate because comprehension of statistics and research methodology is critical to understanding psychology as a science (American Psychological Association [APA], 2007; McGovern, Furumoto, Halpern, Kimble, & McKeachie, 1991). In fact, most psychology programs recommend students take statistics and research methods courses in the second or third year (Perlman & McCann, 1999). Given the conflict between students’ apprehension and advising recommendations, we investigated whether the timing of statistics and research methods courses relates to students’ academic success.

Even though most psychology programs require a methodology course for the major (Messer, Griggs, & Jackson, 1999; Perlman & McCann, 1999), few studies have examined whether the timing of statistics and research methods predicts important academic outcomes (but see Lauer, Rajecki, & Minke, 2006; Scepansky & Carkenord, 2004). Anecdotally, faculty have reported that students take methodology courses late in their academic careers (Rajecki et al., 2005), with some students even waiting until the last semester. Dunn et al. (2010) recommended that students take courses in statistics and research methods after the introductory course in psychology. Following this recommendation, methodology courses could serve as prerequisites for upper-division courses (i.e., junior or senior level) in psychology.

We suspect that the recommendations of Dunn et al. (2010) reflect the logic of many psychology departments’ preferred timeline of courses. In addition, we believe there are several reasons that early enrollment in methodology courses may be beneficial. First, some upper-division courses in psychology implicitly or explicitly expect students to have a rudimentary understanding of statistics and research methodology.
Therefore, understanding of material and performance in upper-division courses might be improved if students have already taken research methodology courses. Second, competence in research methodology is critical to the success of independent research projects (e.g., senior thesis) and research collaborations with faculty. Third, courses in statistics and research methods improve reasoning (VanderStoep & Shaughnessy, 1997), and may enhance critical thinking (Lawson, 1999; McLean & Miller, 2010), especially when care is taken to emphasize how course material applies to “everyday issues” (Mill, Gray, & Mandel, 1994, p. 249). In addition, Penningroth, Despain, and Gray (2007) demonstrated that a 1-credit, first-year course focused on methodology issues (e.g., experimenter bias, correlation vs. causation, etc.) increased students’ psychological critical thinking. Development of critical thinking abilities is especially important because critical thinking skills predict academic performance in psychology courses (Williams, Oliver, Allin, Winn, & Booher, 2003). Therefore, taking statistics and research methods courses early in the psychology major may produce cumulative effects that are positively related to academic performance.

Present Study

Given the significance of methodology courses for advancing the goals of the psychology major and the potential importance that timing of methodology courses might have on academic performance, we examined the transcripts of undergraduate psychology majors to determine whether the timing of statistics and research methods courses matters. Because knowledge of methodology may influence performance in upper-division courses, and methodology courses can foster reasoning skills and critical thinking, we hypothesized that performance in, and enrollment timing of, methodology courses (i.e., statistics and research methods) should predict grade point average (GPA) in upper-division psychology courses. In addition, if comprehension of methodology courses enhances “other spheres of study” (Lauer et al., 2006, p. 24), then methodology course performance and timing may predict students’ accumulated knowledge across a broad range of psychological topics. Therefore, we investigated whether methodology grade and timing of enrollment predicted performance on a test (i.e., Psychology Assessment Test; Pepper, 2006) developed to assess the knowledge of psychology majors enrolled in capstone courses. Finally, in testing these predictions, we controlled for variables (i.e., number of completed psychology courses, relevant American College Test [ACT] scores, and prior GPA) potentially related to methodology performance, timing of enrollment, and academic success.

Method

Participants

We analyzed transcripts of 129 undergraduate psychology majors currently enrolled at or recently graduated (2005–2009) from the University of Wyoming. Because we analyzed transcripts only from students who had finished at least one upper-division psychology course after completing the methodology core (i.e., statistics and research methods), students in the sample had completed a substantial number of psychology courses (\(M = 12.14, SD = 2.43\)). Although transcripts included no demographic information, gender information was collected as part of the Psychology Assessment Test (Pepper, 2006). Of the 65 participants who had taken the Psychology Assessment Test, 55% were women and 45% were men.

Procedure and Materials

Transcripts. Prior to coding, we screened the transcripts of psychology majors to include only those students who had completed the required statistics and research methods courses at the University of Wyoming. In addition, we included transcripts if the student had completed at least one upper-division psychology course (i.e., junior and senior level) after finishing research methods. At the University of Wyoming, statistics is a prerequisite for research methods. Therefore, finishing the research methods course signaled the completion of the methodology requirements. We also included those transcripts for which GPA prior to the onset of the methodology core (i.e., statistics) could be calculated. We used pre-methodology GPA to control for individual variability in prior achievement. Finally, we included only transcripts that contained ACT scores, which we used to account for preexisting differences in cognitive abilities (Coyle & Pillow, 2008; Koenig, Frey, & Detterman, 2008).

To determine the influence of the methodology core on the academic performance of undergraduate psychology majors, we coded for several pieces of information. We included (a) the number of psychology courses completed before statistics, (b) grade in statistics (i.e., \(A = 4, B = 3\), etc.), (c) number of psychology courses taken prior to research methods, (d) grade in research methods, and (e) total number of psychology courses completed. We excluded psychology courses concurrently enrolled with either statistics or research methods from the number of courses completed prior to the specific methodology course (i.e., statistics or research methods). For example, a psychology course taken simultaneously with statistics was not utilized to calculate the number of courses taken before statistics. However, if students completed research methods in a later semester, we used the psychology course in the calculation of the number of courses taken before research methods. We also calculated the cumulative pre-methodology GPA and the GPA from upper-division psychology courses that were taken after the completion of research methods. Grade point average from upper-division psychology classes provided an indicator of upper-division psychology performance after completion of the statistics/research methods core. Finally, we obtained students’ Math and Science Reasoning ACT scores.

Psychology Assessment Test. The Psychology Assessment Test is an online examination developed by the Department of Psychology at the University of Wyoming to assess psychology...
majors’ learning and retention of psychological knowledge (Pepper, 2006). This assessment contains 100 multiple-choice questions designed to evaluate understanding in core areas of psychology, including research methodology, personality, learning, social psychology, developmental psychology, abnormal psychology, cognitive psychology, and biological psychology. The Psychology Assessment Test is a measure of psychological knowledge (Pepper, 2006) administered to students enrolled in 13 upper-division capstone courses. Previous research has shown that senior psychology majors obtain higher scores on the Psychology Assessment Test than first-year psychology majors and seniors who are not majoring in psychology (Pepper, 2006). As our department has only recently implemented the Psychology Assessment Test, only 65 students in our sample had completed the test.

Results

We conceived the courses of statistics and research methods as together fulfilling a methodology core. Therefore, prior to analyses, we calculated the average course grade for the combined statistics and research methods courses. To create a single timing variable for students’ completion of the methodology core, we tallied the number of psychology classes completed prior to statistics and added them to the number of psychology courses completed prior to research methods. For instance, if a student took two psychology classes prior to statistics and four psychology classes prior to research methods, that student received a 6 on the timing of methodology course completion variable. The timing variable is functionally equivalent to the average of the enrollment timing for statistics and research methods. Table 1 shows correlations and descriptive information for all variables.

We used hierarchical multiple regression to explore the importance of statistics/research methods in predicting students’ post–research methods GPA in upper-division psychology courses. Previous research has established the importance of the number of psychology courses completed and math ACT scores in predicting relevant academic outcomes (Frazier & Edmonds, 2002; VanderStoep & Shaughnessy, 1997; Vittengl et al., 2004). In addition, improvement of scientific reasoning is a goal of methodology courses and more broadly the psychology major (APA, 2007; Tomcho et al., 2009), so we believed that individual differences in scientific reasoning, as measured by the Science Reasoning ACT, might relate to both performance in methodology courses and GPA in psychology classes. Finally, previous achievement (i.e., pre-methodology GPA) could influence the timing of enrollment in methodology courses and performance in methodology courses as well as performance in psychology classes. Therefore, in the first block, we entered the number of psychology courses completed, Math ACT score, Science Reasoning ACT score, and pre-methodology GPA. In the second block, we added the average course grade for statistics/research methods and the timing variable.

As expected, the full regression model significantly predicted upper-division psychology GPA, $R^2 = .38$, $F(6, 122) = 12.20, p < .001$. Only pre-methodology GPA ($\beta = .33, p < .001$), average course grade for statistics/research methods ($\beta = .30, p < .01$) and the timing variable ($\beta = -.19, p < .05$) contributed significantly to this model. (See Table 2 for the regression weights.1)

Next, we repeated the same hierarchical multiple regression analysis, but with the Psychology Assessment Test score as the criterion. The sample size for this analysis was smaller ($n = 65$) than the first regression because not all students had completed the Psychology Assessment Test. However, the full regression model significantly predicted performance on the Psychology Assessment Test, $R^2 = .35, F(6, 58) = 6.79, p < .001$. Although timing of methodology course completion was not significant, the number of psychology courses completed ($\beta = .31, p < .01$) and the average grade for statistics/research methods ($\beta = .32, p < .05$) contributed significantly to the model. (See Table 3 for regression weights.) It should be noted however, that the full model was only marginally better than the reduced model at predicting Psychology Assessment Test scores, $R^2 \Delta = .05, F(2, 58) = 2.28, p = .11$.

Discussion

To our knowledge, this study provides the first evidence that timing of statistics/research methods is relevant for the success of psychology majors. Specifically, we demonstrated that the
average course grade from statistics/research methods and the timing of enrollment predicted post–research methods, upper-division psychology GPA, even after controlling for the number of psychology courses completed, relevant ACT scores, and prior achievement (i.e., pre-methodology GPA). In other words, performing well in statistics/research methods and early completion were related to greater academic success in subsequent psychology courses. However, the story was somewhat different when predicting students’ scores on a broad assessment of knowledge in psychology (i.e., Psychology Assessment Test). Average course grade for statistics/research methods, along with the number of psychology courses taken, significantly predicted students’ scores, but timing of enrollment did not.

We are uncertain exactly why timing of enrollment predicted upper-division psychology GPA, but not scores on the Psychology Assessment Test. However, we identified a few possibilities. First, methodology courses may facilitate certain skills (e.g., writing, critical thinking, methodological reasoning, oral presentation) that cumulatively affect GPA, but are less reflected in performance on a multiple-choice test of broad knowledge in psychology. For instance, methodology courses may promote the process of critical thinking (e.g., critically evaluating research literatures and evidence, integrating different literatures in a paper, etc.), whereas the Psychology Assessment Test may emphasize mastery of content knowledge (e.g., what is operant conditioning?). Second, the timing of methodology courses may influence performance in upper-division courses more than lower-division courses. If faculty, explicitly or implicitly, expect students in upper-division psychology courses to understand statistics and research methods, student performance in these courses may rely more on comprehension of methodology than in lower-division courses. In contrast, the Psychology Assessment Test measures knowledge acquired both in lower and upper-division courses. Therefore, the test may be a less sensitive measure of the benefits accumulated after enrollment in statistics/research methods.

Although our research provides preliminary evidence that the timing of statistics/research methods is important, several questions remain. First, we demonstrated that our results were not due to preexisting differences in cognitive ability (i.e., ACT scores) or achievement (i.e., pre-methodology GPA). However, further research is needed to address other important factors (e.g., statistics anxiety, academic self-efficacy, level of self-determination, etc.) that were beyond the scope of this research.

A related limitation is that we did not identify the specific mediator of the enrollment timing and GPA relationship. We assumed that timing of enrollment is important because methodology courses facilitate reasoning, critical thinking, and comprehension of methodology (VanderStoep & Shaughnessy, 1997; Williams et al., 2003). These skills then aid performance in upper-division coursework. Although some evidence exists to support these assumptions, the reason timing of enrollment is important remains an empirical question.

Finally, although most psychology departments recommend methodology courses in the second or third year (Perlman & McCann, 1999), there is variability in the enforcement of enrollment recommendations, the topics taught, type of research experiences, and structure of these courses (Giesbrecht, Sell, Scalifa, Sandals, & Ehlers, 1997; Landrum, 2005; Perlman & McCann, 2005). Therefore, it is unknown if our results are unique to our department or applicable to other academic institutions.

### Implications for Teaching and Advising

If subsequent research can replicate our findings and eliminate potential third variables as the cause of the enrollment timing

---

**Table 2. Summary of Hierarchical Regression Analysis Predicting Upper-division Psychology Grade Point Average (GPA) Post—Research Methods (N = 129)**

<table>
<thead>
<tr>
<th>Predictor</th>
<th>B</th>
<th>SE B</th>
<th>β</th>
<th>R² Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Step 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Psychology Courses Completed</td>
<td>0.03</td>
<td>0.02</td>
<td>.10</td>
<td></td>
</tr>
<tr>
<td>Math ACT</td>
<td>0.01</td>
<td>0.02</td>
<td>.08</td>
<td></td>
</tr>
<tr>
<td>Science Reasoning ACT</td>
<td>0.01</td>
<td>0.02</td>
<td>.06</td>
<td></td>
</tr>
<tr>
<td>Pre-Methodology GPA</td>
<td>0.73</td>
<td>0.11</td>
<td>.51**</td>
<td></td>
</tr>
<tr>
<td><strong>Step 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Psychology Courses Completed</td>
<td>0.03</td>
<td>0.02</td>
<td>.11</td>
<td></td>
</tr>
<tr>
<td>Math ACT</td>
<td>0.002</td>
<td>0.02</td>
<td>.01</td>
<td></td>
</tr>
<tr>
<td>Science Reasoning ACT</td>
<td>0.001</td>
<td>0.02</td>
<td>.003</td>
<td></td>
</tr>
<tr>
<td>Pre-Methodology GPA</td>
<td>0.47</td>
<td>0.13</td>
<td>.33***</td>
<td></td>
</tr>
<tr>
<td>Average GPA (Statistics/Research Methods)</td>
<td>0.35</td>
<td>0.11</td>
<td>.30**</td>
<td></td>
</tr>
<tr>
<td>Timing (Classes Before Statistics/Research Methods)</td>
<td>−0.03</td>
<td>0.01</td>
<td>−.19*p</td>
<td></td>
</tr>
</tbody>
</table>

ACT = American College Test
*p < .05, **p < .01, ***p ≤ .001

**Table 3. Summary of Hierarchical Regression Analysis Predicting Psychology Assessment Test Score (n = 65)**

<table>
<thead>
<tr>
<th>Predictor</th>
<th>B</th>
<th>SE B</th>
<th>β</th>
<th>R² Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Step 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Psychology Courses Completed</td>
<td>0.99</td>
<td>0.39</td>
<td>.27*</td>
<td></td>
</tr>
<tr>
<td>Math ACT</td>
<td>0.47</td>
<td>0.30</td>
<td>.21</td>
<td></td>
</tr>
<tr>
<td>Science Reasoning ACT</td>
<td>0.52</td>
<td>0.27</td>
<td>.26</td>
<td></td>
</tr>
<tr>
<td>Pre-Methodology GPA</td>
<td>6.33</td>
<td>1.80</td>
<td>.38***</td>
<td></td>
</tr>
<tr>
<td><strong>Step 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Psychology Courses Completed</td>
<td>1.15</td>
<td>0.39</td>
<td>.31**</td>
<td></td>
</tr>
<tr>
<td>Math ACT</td>
<td>0.55</td>
<td>0.31</td>
<td>.25</td>
<td></td>
</tr>
<tr>
<td>Science Reasoning ACT</td>
<td>0.37</td>
<td>0.27</td>
<td>.19</td>
<td></td>
</tr>
<tr>
<td>Pre-Methodology GPA</td>
<td>2.91</td>
<td>2.46</td>
<td>.17</td>
<td></td>
</tr>
<tr>
<td>Average GPA (Statistics/Research Methods)</td>
<td>4.63</td>
<td>2.25</td>
<td>.32*</td>
<td></td>
</tr>
<tr>
<td>Timing (Classes Before Statistics/Research Methods)</td>
<td>0.12</td>
<td>0.21</td>
<td>.06</td>
<td></td>
</tr>
</tbody>
</table>

ACT = American College Test
*p < .05, **p < .01, ***p ≤ .001
and GPA relationship, several implications exist for teaching and advising. First, instructors of psychology may already stress the importance of understanding methodology to their students. However, instructors and advisors may also want to encourage early enrollment in methodology courses, and with additional research, some departments may wish to consider a move toward requiring, rather than merely encouraging, early enrollment in methodology courses. Early enrollment may not only benefit students’ comprehension and upper-division GPA, but it may assist in dispelling students’ misconceptions about psychology. For instance, we wonder if delayed enrollment in statistics and research methods may, in part, explain students’ false beliefs regarding psychology (Gardner & Dalsing, 1986) and help to clarify why psychology majors have difficulty seeing the field as a science (Holmes & Beins, 2009).

Ideally, methodology courses should facilitate psychology majors’ critical thinking and ability to think like a scientist (APA, 2007). However, we speculate that postponed enrollment in methodology courses may function to perpetuate preexisting false beliefs, with lengthier delays decreasing the likelihood that students will correct misconceptions or view psychology as a science. For departments that cannot (or do not wish to) require early completion of methodology courses (i.e., statistics and research methods), one solution may be to develop and implement 1-credit, freshman oriented courses that focus on psychological critical thinking (see Penningroth et al., 2007 for specifics). These courses are developed specifically with first-year students in mind and confer many of the benefits (e.g., improvement in critical thinking) associated with methodology courses (Lawson, 1999; Penningroth et al., 2007; VanderStoep & Shaughnessy, 1997).

As a second implication of our research, programs in psychology may find it informative to conduct transcript analyses of recent graduates to determine if students are postponing methodology enrollment. Even if departments explicitly recommend or “require” early enrollment (Perlman & McCann, 1999), we surmise that if anxiety or trepidation regarding methodology courses is strong enough (Onwuegbuzie & Wilson, 2003; Rajecck, et al., 2005), inventive college students will find ways around early enrollment unless it is an enforced requirement. Transcript analysis may also identify factors that inadvertently contribute to late enrollment in methodology courses (e.g., transferring from a 2-year to a 4-year college, changing majors, etc.).

Finally, although not examined in this study, we believe that emphasizing methodology across the curriculum could facilitate students’ performance in upper-division courses. Even though early enrollment in methodology courses may have positive implications, too many students delay taking these classes (Perlman & McCann, 2005). By incorporating methodology and research across multiple lower-division classes, faculty can ensure that students learn methodological concepts early in their academic careers, regardless of when they choose to enroll in statistics or research methods.

Conclusion
Psychology students may approach statistics and research methods courses with anxiety and disdain (Onwuegbuzie & Wilson, 2003; Rajecck, et al., 2005), but our research suggests that early enrollment and performance in statistics/research methods predict success in psychology majors’ academic careers. Statistics/research methods may indeed be “painful” for psychology majors, but these courses are also fundamental for promoting comprehension of psychology as a science (APA, 2007; McGovern et al., 1991) and relate to performance in upper-division psychology courses.

Acknowledgments
The authors thank Galen Hiatt and Danielle Quinn for their assistance with transcript coding.

Declaration of Conflicting Interests
The authors declared no potential conflicts of interests with respect to the authorship and/or publication of this article.

Financial Disclosure/Funding
The authors disclosed receipt of the following financial support for the research and/or authorship of this article: This research was supported in part by a grant from the Ellbogen Center for Teaching and Learning at the University of Wyoming.

Note
1. We also examined the possibility that conscientiousness could account for the effect of the timing variable on psychology GPA. Conscientiousness may relate to enrollment timing in methodology courses, as well as academic effort and performance (Trautwein, Lüdtke, Roberts, Schnyder, & Niggli, 2009). Therefore, we acquired Big Five Inventory (John, Donahue, & Kentle, 1991) data on conscientiousness from the fifth author and were able to match these scores to 26 transcripts in our sample. The Conscientiousness scale demonstrated adequate reliability (α = .78). When controlling for conscientiousness, in addition to the previous control variables (i.e., number of psychology courses completed, ACT scores, and pre-methodology GPA), the full model did not reach significance, R² = .15, F(7, 18) = 1.64, p > .05, but average grade in statistics/research methods remained a significant predictor (β = .85, p < .05). Although the beta weight for timing did not reach significance, it was similar (β = -.16) to the beta weight without conscientiousness in the model (i.e., β = -.19). Despite the fact that the beta weights for methodology grade and timing were somewhat similar when controlling for conscientiousness, this analysis was underpowered (power < .80); therefore future research needs to examine conscientiousness with a larger sample. We were unable to conduct a regression analysis with the Psychology Assessment Test as the criterion because of the small overlapping sample (n = 13) with conscientiousness data.

References


**Bios**

**Scott Freng** is an academic professional lecturer in the Department of Psychology at the University of Wyoming. He received his PhD (Social Psychology) from the University of Nebraska-Lincoln in 2001.

**David Webber** is currently a graduate student in the Department of Psychology at the University of Alberta. He received his MS (Psychology) from the University of Wyoming in 2009.

**Jamin Blatter** is a graduate student studying Experimental Psychology at the University of Colorado at Colorado Springs. She received her BA in Psychology from the University of Wyoming in 2009.

**Ashley Wing** is a student at the University of Wyoming’s College of Law. She received her BA from the University of Wyoming in 2010.

**Dr. Walter Scott** is an associate professor who studies basic cognitive and affective processes in personality functioning and depression.
Increasing Academic Self-Efficacy in Statistics With a Live Vicarious Experience Presentation

Robert A. Bartsch¹, Kim A. Case¹, and Heather Meerman¹

Abstract
This study investigated the effect of a vicarious experience on the academic self-efficacy of graduate students enrolled in a statistics and research methods course. Participants (N = 39) completed a self-efficacy scale during the first two meetings of the course. Two weeks later, a portion of these students participated in a randomly assigned intervention to increase statistics self-efficacy. In the experimental condition, a former statistics student came to the class and explained her own math anxieties and outlined the behaviors that led to her personal success in the same course. Comparison students wrote about the characteristics of a successful student in the course without the experience of a peer model presentation. Analysis of pre- and postintervention academic self-efficacy indicated students in the peer model group showed a significant increase compared to the writing group. We discuss the potential of using live vicarious experience presentations to increase self-efficacy in psychology statistics courses.

Keywords
self-efficacy, teaching statistics, vicarious experience

Psychology instructors consistently face student anxieties, fears, and cognitive obstacles associated with taking statistics courses. Low self-efficacy with regard to math and statistics often affects performance (Bong & Skaalvik, 2003), leading to many studies exploring the links between self-efficacy and success in mathematics (Finney & Schraw, 2003; Friedel, Cortina, Turner, & Midgley, 2010; Harlow, Burkholder, & Morrow, 2006; Kennett, Young, & Catanzaro, 2009). Undergraduate and graduate psychology courses in statistics and research methods require innovative approaches to help students strengthen their own personal statistics self-efficacy during early stages of the course. Schunk (1991) defined academic self-efficacy as students’ beliefs in their own abilities to successfully perform a set of given academic tasks, such as passing a course. A multitude of studies identify self-efficacy as the strongest predictor of academic performance level (Bandura, Barbaranelli, Caprara, & Pastorelli, 1996; Ferla, Valcke, & Cai, 2009; Hackett, Betz, Casas, & Rocha-Singh, 1992; Lane & Lane, 2001; Lane, Lane, & Kyprianou, 2004; Lent, Brown, & Gore, 1997; Lopez & Lent, 1992; Multon, Brown, & Lent, 1991). Self-efficacy in academic contexts also predicts grade goals, cognitive strategy use, self-regulation, and intrinsic interest (Bong & Skaalvik, 2003).

Research suggests that the modeling of success through vicarious experience (Bandura, 2008; Lane et al., 2004) may have the most instant and direct influence on self-efficacy (Gist & Mitchell, 1992; Luzzo, Hasper, Albert, Bibby, & Martinelli, 1999). During a vicarious experience, a model informs and motivates by providing information about specific behaviors and strategies that led to personal success in that particular situation (Schunk, 2003). Witnessing the successful performance by a model supports the belief in one’s own ability to perform the task, such as passing a statistics course, successfully. Prieto and Meyers (1999) included vicarious learning in their assessment of teaching assistant training and supervision, but the variable was not specifically isolated as a predictor of self-efficacy. Harlow et al. (2006) investigated the impact of several pedagogical strategies, including semester-long peer mentoring, on self-efficacy within a quantitative psychology course. Although students’ reported statistics self-efficacy increased by the end of the term, instructors need efficient ways to increase academic self-efficacy during very early stages of these courses. Luzzo et al. (1999) examined vicarious experience by showing undergraduate students a 15 min video presentation of two graduate students describing their personal experiences with math and science. The video alone failed to significantly influence math and science self-efficacy among student viewers and the researchers recommended live models for intervention success. Although this

¹ University of Houston–Clear Lake, Division of Human Sciences, Houston, Texas, USA

Corresponding Author:
Robert A. Bartsch, University of Houston–Clear Lake, Box 15, U of Houston Clear Lake, 2700 Bay Area Blvd, Houston, TX 77058, USA
Email: bartsch@uhcl.edu
particular approach utilized vicarious learning, the models represented a higher level of education than the viewers, which may have weakened the effectiveness of the intervention. In general, greater similarity between the observer and the model contributes to greater persuasiveness of the model (Bandura, 1997, 2008; Schunk, 1987).

In their influential paper distinguishing academic self-concept from academic self-efficacy, Bong and Skaalvik (2003) called for additional research on instructional approaches that will increase student self-efficacy in the context of specific courses. Psychology instructors may benefit from the use of peer models to enhance self-efficacy among students, especially within courses that produce greater levels of anxiety among students. The current study measured the impact near the beginning of the semester of a live vicarious experience intervention, a presentation by a former student that successfully completed the course, on students’ perceived academic self-efficacy with regard to successful completion of a statistics course. We expected students witnessing a live presentation from a similar peer model to report a significant increase in academic self-efficacy compared to those with no peer model. The findings will aid teachers of psychology aiming to increase student perceptions of self-efficacy with respect to academic performance in a particular course.

Method

Participants

Participants included 64 students from three sections of a master’s-level research design and statistics course at a medium-sized state university. The sample consisted of 54 females and 10 males. The groups’ average age was 28 years (range: 21–63). Students identified as White (61%), African American (6%), Hispanic/Latino (19%), or other (14%). Students did not receive any extra credit or other incentive for their participation.

Instruments

The self-efficacy scale (Cronbach’s α = .94) used in this study was developed based on a measure from Lane and Lane (2001). Our scale measured the confidence on competencies needed to achieve success in the research design and statistics course. We asked participants to answer eight self-efficacy questions on a 10-point Likert-type scale ranging from (1) no confidence to (10) total confidence. Sample items from the scale included: “You have sufficient math ability to meet the demands of the course” and “You can manage your time to meet the demands of the course.”

Procedure

At the beginning of the semester, we asked all students in the class to complete informed consent forms along with the self-efficacy scale. Participants were not told about any future measurements. Two weeks after this pretest, we randomly assigned students to experimental and control group interventions and placed these groups in separate rooms. A previous research and statistics course graduate student made a 10 min presentation to the experimental group. The female peer model presenter, close to the average participant age, described her performance level as similar to the majority of students enrolled in the course. The presenter discussed her experiences in the course, how she managed her time, her study methods, and how she managed stress related to the course. She also outlined her eventual success in the course. To make the presentation more realistic, the student did not have a script but rather an outline of what she would say. Questions or comments were not allowed after her vicarious presentation. After the presentation, students completed a posttest self-efficacy scale identical to the pretest. It is possible that some participants knew or knew of the student, but we believe most did not.

In a separate room, control group participants were told to imagine a successful student in a research and statistics course. While reflecting on this successful peer, control group participants wrote for up to 10 min about this student’s time management, study habits and methods, and stress management skills that lead to success in the course. After receiving the instructions, students wrote about the successful student in whatever form they wanted. Following the reflective writing exercise, students completed the posttest self-efficacy scale.

Results

We conducted a 2 × 2 factorial ANOVA to determine whether testing time (pre-intervention vs. postintervention) and group (control vs. experimental) had an effect on participants’ overall self-efficacy ratings. We used the pre-intervention ratings to control for initial levels of self-efficacy as well and to determine any changes in self-efficacy. As expected, analyses indicated no significant main effect of testing time, $F(1, 37) = 0.07, ns$, or group type, $F(1, 37) = 0.11, ns$. In addition, the ANOVA revealed a significant Testing Time × Group interaction, $F(1, 37) = 9.38, p = .004$, partial $\eta^2 = .20$. A marginally significant increase in academic self-efficacy ratings occurred for students in the experimental group from pre- ($M = 6.66$) to postintervention ($M = 7.07), t(19) = −2.01, p = .06, d = 0.45.$ Control group participants exhibited a significant decrease in self-efficacy from pre- ($M = 6.96$) to postintervention ($M = 6.48), t(18) = 2.32, p = .03, d = 0.53.$

Discussion

The findings suggest that students may improve self-efficacy by hearing about previous course experiences from former successful students in person. In addition, a presentation from an average student, as opposed to the “star pupil,” may be especially useful to increase the effectiveness because current students recognize similarities between themselves and the model. In fact, to counter low self-efficacy in students, it may be important for the model to discuss experiencing low academic self-efficacy upon first entering the course. In the control condition, student reflections and writing about an
imagined successful student failed to improve self-efficacy and even indicated a possible decreased self-efficacy among participants. When imagining a successful student, participants may think of somebody quite different from themselves and consequently assume that the fictitious model possessed unique qualities necessary for success. The decreased self-efficacy among control students may also be influenced by learning more about the course requirements and expectations in the first two weeks of the course, as well as initial encounters with the textbook and statistics assignments.

In this study, providing a vicarious experience from a successful model, who students perceive as similar to themselves, led to an increase in self-efficacy. Although this study focused on master’s-level students in a research and statistics course, the live vicarious experience peer model presentation may also enhance academic self-efficacy in both undergraduate and graduate psychology statistics and research methods courses. However, we also have tentative evidence that a poor presentation by a vicarious model may reduce academic self-efficacy. Furthermore, the medium itself (video vs. live) may influence student levels of academic self-efficacy. The live interaction could provide a more realistic or authentic presentation, leading to greater persuasion. Viewing a video (Luzzo et al., 1999) may feel more distant and more difficult to connect with personally. In contrast, a live presentation provides students with more opportunities to make connections and find similarities between themselves and the former student. Future research might explore the video versus live presentation methods and measure perceptions of persuasiveness. Additionally, further research could examine the role that perceived similarity to the peer model plays in improving self-efficacy. Several aspects of the study design and sample selection limit broad application of the findings. Given that students completed the pre- and posttest surveys in class, self-report bias may have influenced the responses. Although none of the researchers collecting the data taught the courses involved in data collection, results may still be affected by demand characteristics; however, given that both groups had an intervention, it is unclear why the students in one intervention would be affected more by demand characteristics than another. Furthermore, it is unclear whether this finding is limited to anxiety-producing courses such as statistics or if this finding can be broadened to other areas of academic self-efficacy in other courses or even degree attainment.

Declaration of Conflicting Interests
The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The authors received no financial support for the research, authorship, and/or publication of this article.

Note
1. After one of three sections and before results were tabulated, the peer model informed a researcher that she did not feel she did a good job of expressing herself in her presentation. She stated she felt nervous and did not feel she was a good model. Given this situation, we removed that section (N = 25) from the main analysis. If all three sessions are combined in one analysis with session as a factor, the Testing Time × Group interaction is no longer significant, F(1, 58) = 1.19, ns, but there is a Testing Time × Group × Section triple interaction, F(2, 58) = 4.90, p = .01, indicating that the differences between conditions were different between the three sections. With all participants, the simple effect of intervention in the experimental condition was also marginally significant, t(32) = –1.80, p = .08, d = 0.31. In the third section, in which the peer model did not perform well, students exhibited a different pattern of results. There was an increase in the control condition from the pretest (M = 6.70) to the posttest (M = 7.22) and no change in the experimental condition from the pretest (M = 7.18) to the posttest (M = 7.21).

References
importance of learned resourcefulness. *Educational Psychology*, 29, 815-830.
Statistical Training in Psychology: A National Survey and Commentary on Undergraduate Programs
James Friedrich, Evelyn Buday and David Kerr
Teaching of Psychology 2000 27: 248
DOI: 10.1207/S15328023TOP2704_02
The online version of this article can be found at:
http://top.sagepub.com/content/27/4/248

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Society for the Teaching of Psychology

Additional services and information for Teaching of Psychology can be found at:

Email Alerts: http://top.sagepub.com/cgi/alerts
Subscriptions: http://top.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Oct 1, 2000
What is This?
We surveyed a national sample of U.S. undergraduate psychology programs regarding the structure and content of statistical training. Results revealed considerable diversity in approaches and offerings. Contemporary trends in data analysis (e.g., power and effect size analysis, confidence interval estimation, general linear model approaches) as well as measurement issues appear to receive relatively little attention in the core sequence. Respondents tended to view such topics as appropriate for more advanced courses, which they said were infrequently required. We discuss options for addressing issues of course sequencing, content focus, and advanced-level offerings in the major.

Psychologists generally view statistical training for undergraduate majors as an essential component of the curriculum (Brewer et al., 1993; McGovern, Furumoto, Halpern, Kimble, & McKeachie, 1991), with most U.S. schools listing it as a requirement for the major (Messer, Griggs, & Jackson, 1999; Perlman & McCann, 1999a, 1999b). What is unclear from these discussions and reviews of catalog descriptions is the actual content of this instruction. Such detailed curricular information has been available only for statistical training at the doctoral level (Aiken, West, Sechrest, & Reno, 1990). The data we report here explore in similar detail the content of undergraduate statistics instruction.

There are several reasons to explore the statistical curriculum more closely. Perhaps most important, it is unclear whether recent critiques and advances in statistical analysis are making their way into undergraduate instruction. We are not referring to advanced statistical techniques that are clearly the province of graduate education. Rather, we have in mind several general "themes" that cut across even more elementary aspects of measurement and data analysis. For example, exploratory data analysis (EDA) approaches, effect size computation, statistical power considerations, and the relative merits of null hypothesis testing versus confidence interval estimation (Behrens, 1997; Cohen, 1992, 1994; Frick, 1998; Harris, 1997; Loftus, 1996; Rosenthal & Rosnow, 1991; Wainer, 1999) are issues that have generated considerable discussion among psychologists and data analysts in recent years. Many of these concerns were nicely highlighted in an interview with Chris Spatz (Dillon, 1999) in the pages of this journal. The importance of these issues is reflected in the American Psychological Association’s (APA) appointment of a Task Force on Statistical Inference. Aspects of its recently published guidelines (Wilkinson & Task Force on Statistical Inference, 1999) will likely become codified in subsequent revisions of the Publication Manual of the American Psychological Association (1994).

Undergraduate instructors’ willingness to address these issues in foundation courses may be critical to eventually incorporating such approaches into mainstream practice. In addition, instructors’ attention to certain broader themes such as the general linear model approach to integrating ANOVA and regression (e.g., Aron & Aron, 1999) and model comparison approaches for testing theoretical claims (e.g., Judd, McClelland, & Cuihane, 1995; Lockhart, 1998) might also help establish greater coherence and flexibility to undergraduates’ statistical training.

From a pragmatic point of view, one might also ask the basic question of whether current coverage meets the needs of majors in a rapidly progressing empirical discipline. For example, undergraduate instruction in psychology has traditionally involved some exposure to professional journal literature, but the kinds of statistics discussed extensively in empirical articles tend to fall outside the domain of most introductory courses (cf. Sherman, Buddie, Dragan, End, & Finney, 1999). Similarly, many undergraduate theses and senior research projects draw on designs and techniques that instructors may be unable to cover adequately in the typical semester course ending with an introduction to one-way ANOVA.

The report from APA’s Task Force on Statistical Inference (Wilkinson et al., 1999) did encourage researchers to use the “minimally sufficient analysis” needed to address one’s research question effectively. However, even if such a move toward “less is more” reduces the complexity of statistical reporting in some journal articles, the use of basic techniques—reliability estimates, multiple regression, and focused contrasts, for example—would continue to pose challenges for undergraduates grappling with research articles and senior projects. It is tempting to think that more rigorous statistical training is a concern only for students headed to graduate programs in psychology. However, 248
tative and analytical skills are also important strengths that undergraduate psychology majors might bring to the marketplace (Clough, 1993). Psychology’s increasing sophistication may simply carry with it a need for more extensive methodological and statistical training in the major.

To obtain an accurate picture of current instructional practices, we conducted a direct mail survey of a large random sample of U.S. undergraduate programs, modeling our strategy after Aiken et al.’s (1990) survey of psychology doctoral programs. Specifically, we explored for both introductory and advanced statistics courses (a) the ways in which statistics and measurement instruction is being incorporated into the curriculum, (b) the extent to which current statistical themes are being integrated into instruction, and (c) the nature of the statistical techniques covered. Given the variability that might be associated with differences in student selectivity and faculty background, we followed Aiken et al.’s framework in sampling so as to permit explicit comparisons between an “elite” and a general sample of programs.

Method

Sample

We obtained a complete listing of U.S. institutions offering bachelor’s degrees, along with admission test scores, undergraduate enrollments, and other demographic information from the annual U.S. News & World Report (1997) directory of colleges and universities. We excluded schools devoted exclusively to programs in business, engineering, fine and performing arts, and military service, leaving 1,319 accredited institutions in the population from which we sampled.

Our design called for a sizable sample from the elite end of the distribution. Because of our focus on undergraduate instruction, however, we did not rely on reputational or publication rankings of specific psychology programs (e.g., Howard, Cole, & Maxwell, 1987), as these rankings tend to overrepresent schools with graduate faculty. Instead, we selected all institutions included in U.S. News & World Report’s (1997) top-50 ranked national universities and top-40 ranked national liberal arts colleges. In addition, we drew a simple random sample of 315 from the remaining colleges and universities, bringing the total number of schools eventually contacted to 405. The elite and general samples differed substantially in terms of such indicators as student median SAT (Scholastic Assessment Test) scores and high school ranks as well as in terms of resources and reputational rankings (for selection criteria, see U.S. News & World Report, 1997). Thus, without judging what other worthy schools we might have included, our elite group appears to reflect high consensus choices.

Psychology department chairs at the sampled institutions received a multipage survey. In a cover letter describing the nature of the project, we asked the chairs to provide basic demographic and curricular information and then to direct the questionnaire to those faculty best able to comment on instruction in the undergraduate statistical courses. Two successive followups for nonresponders produced a final total of 255 returned surveys (63%). Some responses were incomplete; the analyses reported here reflect slightly varying sample sizes as a result. A comparison of responding and nonresponding institutions revealed no significant differences in entrance test scores, high school class ranks, undergraduate enrollments, presence of graduate programs, or choice to have statistics taught inside versus outside of the department. In addition, the general and the elite samples replied at comparable levels (62% and 68%, respectively).

Survey

The first page of the questionnaire requested general background information on the program and department, followed by a section on statistical and methodological requirements for the undergraduate major. Specifically, respondents indicated whether the department offered elective or required courses in each of the following categories: (a) general research methods, (b) thematic methods (e.g., methods in social psychology), (c) introductory statistics, (d) integrated statistics and methods, (e) advanced statistics specifically for undergraduates, (f) graduate-level statistics open to undergraduates, and (g) psychological tests and measurements. They also indicated whether the introductory statistics course or its equivalent was taught within the psychology department, how the statistics and methods courses were sequenced if taught independently, and the extent to which introductory statistics served as a prerequisite for other nonmethods courses within the department.

The second page contained a listing of various basic and intermediate statistical procedures (e.g., independent groups $t$ test, simple regression; for a full description, see Table 3). We asked respondents to categorize coverage within one of four levels, labeled to correspond roughly to the number of class hours devoted to each topic, ranging from 0 (no coverage or very brief mention), 1 (modest but non-trivial coverage; e.g., a single class period or significant portion of one devoted to the topic), 2 (substantial coverage; e.g., up to a semester week—3 class hr—devoted to the topic; total time might, however, be spread over the course of a term), to 3 (extensive coverage; e.g., more than a semester week devoted to the topic area; would qualify as a central focus of the course).

Respondents first estimated “the degree of coverage each receives in the course or courses your psychology majors use to obtain their initial, intensive exposure to statistics.” They then returned to the list and estimated “the degree of coverage each receives in the course your psychology majors would likely use to obtain their next, more advanced level of statistical training.” If such a course was not currently available to undergraduate majors, respondents indicated the coverage they would desire if they could offer it. It is important to note that advanced statistical training takes place in a variety of contexts (e.g., general and thematic methods courses having a statistics prerequisite, dedicated courses in advanced statistics). Thus, we used fairly general phrasing in asking respondents to characterize advanced coursework. Responses for advanced level instruction therefore reflected a more heterogeneous array of existing and anticipated courses than was true for introductory instruction.

The third page of the survey presented a series of themes in statistical instruction relevant to a broad range of statistical tests, including such topics as power analysis and confidence...
interval computation (for a complete listing, see Table 2). The instructions noted that these issues cut across various techniques and that estimates of coverage might overlap with estimates on the preceding page (e.g., much time spent on t tests might have been devoted to power analysis and confidence intervals). We asked respondents to include such time in these new estimates, even if already noted implicitly in the preceding questions. They again replied on the 0 to 3 scale, separately for the introductory course and the advanced course. The final page of the survey solicited information regarding statistical software usage, course goals, and instructor specialties, analysis of which we omit here in the interest of space.

Results

Because of the primarily descriptive goals of the survey and the large number of interesting but nontheory driven comparisons, we have followed the format of previous surveys of this sort (e.g., Aiken et al., 1990) by avoiding the reporting of statistical significance tests. We have adopted a conservative stance, however, in identifying differences to highlight and discuss. In terms of stratification variables, we looked beyond the elite–general sample distinction to also compare programs requiring an integrated statistics and methods sequence to those requiring separate statistics and methods courses as well as programs teaching introductory statistics within psychology to those requiring a course outside the department. In the interest of space, the tables reflect only the elite versus general sample breakdowns central to the sample, with interesting findings along these other dimensions described in the text.

General Curricular Structure

We asked respondents to specify their current statistical and methodological requirements as part of the survey, retrieving limited descriptive information from institutional catalogs and Web pages only for nonresponders. Of the 405 schools sampled, 25 did not appear to offer a psychology major. We obtained relatively complete information for 373 of the remaining 380 schools. Table 1 reports the percentage of schools from the elite and general samples that required, offered as an elective, or did not offer the listed statistics and methodological courses. Considerable exposure to methodology and statistics may also occur in more topically oriented content classes (e.g., in a lab attached to a cognition course); the courses included in the survey merely reflect those in which these topics are the primary focus.

Although organized somewhat differently for the purposes of this study, our results are generally similar to those reported in recent catalog-based surveys (Messer et al., 1999; Perlman & McCann, 1999a, 1999b). Table 1 summarizes the responses separately for the elite and general samples; responses for the combined sample have been omitted to streamline the table but can be readily inferred from the tabulated values. To clarify general trends, we note selected results for the combined sample as well as certain supplementary analyses in the following paragraphs.

Virtually all of the programs—89% of the combined general-plus-elite sample—required a methodology course of some sort (general, thematic, or integrated). Research methods courses tended to be taught as general methodology courses rather than as thematic courses tied to particular content areas. A statistics course of some sort was required by 93% of the combined sample, most commonly offered as a stand-alone course (86%). However, a significant minority of schools (26%) reported offering integrated statistics and methods courses. (The combined total exceeded 100% because some programs reported offering both segregated and integrated courses.) Only 41% of the programs offered some kind of advanced statistics course open to undergraduates and just 6% actually required such a course of majors. Because some schools incorporate more advanced techniques within an integrated statistics and methods sequence, these latter figures are likely conservative with respect to opportunities for instruction beyond introductory statistics.

These patterns differed somewhat between the elite schools and the general sample (see Table 1). Specifically, the elite schools were less likely than the general sample to require a general methods course and were more inclined to offer thematic methods courses. The elite schools were also more likely to offer an advanced statistics course, although the greater availability of graduate courses open to undergraduates appears attributable at least in part to the greater

### Table 1. Percentages of General Sample and Elite Sample Institutions Offering Various Statistics and Methodology Courses

<table>
<thead>
<tr>
<th>Course</th>
<th>Not Offered</th>
<th>Elective</th>
<th>Required</th>
<th>With Lab (% of Elective Plus Required)</th>
</tr>
</thead>
<tbody>
<tr>
<td>General research methods</td>
<td>17.8</td>
<td>14.0</td>
<td>7.8</td>
<td>37.3</td>
</tr>
<tr>
<td>Thematic methods courses</td>
<td>72.8</td>
<td>80.8</td>
<td>80.9</td>
<td>30.8</td>
</tr>
<tr>
<td>Introductory statistics</td>
<td>75.2</td>
<td>72.6</td>
<td>72.9</td>
<td>38.0</td>
</tr>
<tr>
<td>Advanced statistics (undergraduate course)</td>
<td>77.7</td>
<td>88.5</td>
<td>6.3</td>
<td>12.5</td>
</tr>
<tr>
<td>Advanced statistics (graduate course open to undergraduates)</td>
<td>50.0</td>
<td>61.0</td>
<td>1.2</td>
<td>20.0</td>
</tr>
<tr>
<td>Psychological tests and measurements</td>
<td>23.8</td>
<td>11.2</td>
<td>19.2</td>
<td>6.0</td>
</tr>
</tbody>
</table>

Note. General sample, n = 286 to 287. Elite sample, n = 85 to 87.
percentage of graduate programs among the elite sample. Finally, the elite schools were consistently more likely to offer the various statistics and methodology courses with accompanying labs. It is unclear, however, whether this finding reflects a philosophical difference or other factors such as more specialized faculty interests or advantages in facilities and staffing afforded by their resources and teaching loads.

Despite strong recommendations that a tests and measurement course be required as part of all variations of the psychology major (Brewer et al., 1993; McGovern et al., 1991), this area appears to be neglected at the undergraduate level much as it has been at the graduate and professional levels (Aiken et al., 1990; Lambert, 1991; Meier, 1993). Only 13% of the combined sample required this course, and 30% reported that they did not offer an undergraduate course. Surprisingly, this neglect of measurement coursework was particularly apparent in the elite sample, where only one school reported requiring the course, and 50% did not offer the course. It is possible that some of these issues are now receiving substantial attention in statistics and methods courses, but data presented in subsequent sections suggest that this is not the case.

Given the survey’s focus, we attempted to further clarify the structure and sequencing of statistics and methods requirements. Patterns for the elite schools and the general sample were quite comparable, so the following percentages of schools adopting each approach are based on the combined sample. Among all the schools for which data were available (n = 373), 14% did not require a separate introductory statistics course but instead required one or more terms of an integrated statistics and methods course. Another 7% either did not offer statistics courses (separately or integrated) or offered them only as electives. However, the vast majority of programs (79%) required a separate, introductory statistics course, most typically in conjunction with a general or thematic methods course or in some cases an additional integrated statistics and methods course. Among schools in this subset that also required a separate, general, thematic, or integrated methods course (n = 280), the overwhelming preference for sequencing was to require that students take the statistics course first (82%). Only 6% required the methods course to be taken prior to the statistics course, with another 11% allowing statistics and methods to be taken in any order, and 1% requiring them to be taken concurrently. Finally, among schools requiring either a separate introductory statistics course or an integrated statistics and methods course, 28% reported the course as being taught outside the psychology department.

We also asked respondents to estimate the extent to which introductory statistics or its equivalent served as a prerequisite to other major courses (exclusive of methods). Given the anticipated difficulty in comparing numbers of courses across schools differing in department size and offerings, we asked respondents (n = 249) to answer on a subjective ordinal scale. Specifically, 22% indicated that introductory statistics or its equivalent was not a prerequisite for any intermediate or upper level psychology courses, 45% indicated that it was a prerequisite for “only a very few” such courses, 18% characterized it as a prerequisite for “numerous” such courses, and 15% characterized it as a prerequisite for “most” intermediate and upper level courses. Elite schools, as well as those adopting an integrated statistics and methods approach, indicated a somewhat greater prerequisite role for statistical training. In general, however, all breakdowns indicated that coursework incorporating basic statistics does not serve a primary role as a prerequisite to other department offerings.

Content of Statistical Instruction

Statistical themes. The focus of the remainder of the survey was to explore in detail the material covered in the introductory and next-more-advanced statistics courses. Although the division between statistical themes (Table 2) and more specific statistical techniques (Table 3) represents to some degree a convenience, it also captures important distinctions central to recent debates over appropriate statistical practice and instruction (Wilkinson et al., 1999).

As shown in Table 2, survey respondents indicated how much emphasis each listed theme received over the course of the term. Although schools differed with respect to semester versus quarter calendars, we tried to maintain a common metric by anchoring the scale to the number of in-class hours devoted to the topic. An inspection of Table 2 reveals that the absolute levels of emphasis reported for the various themes was quite low. For example, more than half of the combined sample reported that for the introductory course, a hour or less of class time is devoted to each of confidence interval estimation, power analysis, effect size estimation, graphical analysis of data, and APA style for statistical reporting—all themes that lie within the scope and difficulty level of an introductory course and are treated in numerous introductory texts. Ninety percent of respondents reported giving no more than a class hour to discussion of the equivalence of ANOVA and regression in terms of a general linear model approach, with similar figures for most of the remaining themes.

In terms of the stratification variables, the elite sample reported that their introductory statistics offering (or its equivalent) placed somewhat greater emphasis on issues of statistical power, effect size estimation, general linear model approaches, and assessment of assumption violations than did the general sample. Introductory courses taught outside the psychology department, as well as those taught as separate rather than as integrated (methods) courses, gave greater attention to confidence intervals; introductory courses taught within psychology departments placed greater emphasis on power and APA reporting style. Finally, those courses taught as integrated rather than as separate courses gave somewhat greater attention to measurement issues, statistical control, assessing assumption violations, and APA reporting style.

It is certainly appropriate that some of the more advanced themes receive little treatment or mention in an introductory course. As is clear from Table 2, respondents viewed greater coverage as appropriate in courses beyond the introductory level. Although confidence interval coverage appeared no more a priority in the advanced course than in the introductory course, respondents indicated greater emphasis on all the other themes. In absolute terms, a majority of respondents judged power analysis, effect size estimation, the general linear model, statistical control, and assessment of assumption viola-
tions to merit substantial attention in the advanced-level course, with the elite sample reporting greater emphasis than the general sample on the latter two. Such themes as confidence interval estimation, measurement issues, and model comparison approaches, however, continued to receive weak emphasis in terms of class time allocation.

Statistical techniques. Respondents used a similar 4-point scale (again corresponding roughly to the number of dedicated class hours) to indicate their emphasis on various analytical procedures and tests (see Table 3). For the combined sample, the introductory curriculum tended to emphasize correlation, between-subjects t tests, and one-way ANOVA (although not likely as part of a general linear model approach, given the results in Table 2). Simple regression got significant but lesser coverage than these other topics, with probability and contingency table analysis receiving still less, and nonparametrics joining topics beyond one-way ANOVA in receiving little attention.

For the introductory course or its equivalent, elite institutions differed little from the general sample except for a reported tendency to devote more attention to between-subjects factorial designs along with focused contrasts and comparisons. This difference may be due to the elite institutions progressing at a bit faster rate and getting further along in a typical text’s chapter structure. In terms of other stratification variables, teaching the course inside as opposed to outside the department was associated with a greater reported emphasis on virtually all topics up through ANOVA, with the exception of probability. This pattern reflects psychology courses’ greater emphasis on data analytic procedures and lesser focus on probability and perhaps other more mathematical or theoretical topics not included in the survey list. Finally, teaching the introductory course as part of an integrated sequence was associated with relatively less emphasis on probability and a greater focus on topics related specifically to ANOVA, including factorial and repeated measures designs as well as post hoc tests and planned contrasts.

Table 2. Percentage of Respondents Claiming Each Level of Coverage for Statistical Themes

<table>
<thead>
<tr>
<th>Statistical Themes</th>
<th>Introductory Course (Coverage Level)</th>
<th>Advanced Course (Coverage Level)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Confidence interval estimation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>General</td>
<td>15</td>
<td>36</td>
</tr>
<tr>
<td>Elite</td>
<td>11</td>
<td>45</td>
</tr>
<tr>
<td>Statistical power</td>
<td></td>
<td></td>
</tr>
<tr>
<td>General</td>
<td>21</td>
<td>53</td>
</tr>
<tr>
<td>Elite</td>
<td>6</td>
<td>39</td>
</tr>
<tr>
<td>Effect size estimation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>General</td>
<td>43</td>
<td>45</td>
</tr>
<tr>
<td>Elite</td>
<td>35</td>
<td>26</td>
</tr>
<tr>
<td>Equivalence of ANOVA and regression (the general linear model)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>General</td>
<td>69</td>
<td>21</td>
</tr>
<tr>
<td>Elite</td>
<td>52</td>
<td>35</td>
</tr>
<tr>
<td>Psychometrics (e.g., classical test theory, internal consistency estimates, reliability and attenuation of effect sizes)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>General</td>
<td>64</td>
<td>25</td>
</tr>
<tr>
<td>Elite</td>
<td>56</td>
<td>33</td>
</tr>
<tr>
<td>Statistical control through “partial effects” (e.g., partial correlation and regression coefficients, ANCOVA)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>General</td>
<td>81</td>
<td>16</td>
</tr>
<tr>
<td>Elite</td>
<td>78</td>
<td>11</td>
</tr>
<tr>
<td>Assessing violations of model assumptions (including outlier analysis)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>General</td>
<td>51</td>
<td>39</td>
</tr>
<tr>
<td>Elite</td>
<td>29</td>
<td>49</td>
</tr>
<tr>
<td>Graphical analysis of data (e.g., box and whisker, residual plots)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>General</td>
<td>26</td>
<td>37</td>
</tr>
<tr>
<td>Elite</td>
<td>27</td>
<td>35</td>
</tr>
<tr>
<td>Model comparison approaches (comparing compact versus expanded models in terms of incremental reduction in error)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>General</td>
<td>87</td>
<td>9</td>
</tr>
<tr>
<td>Elite</td>
<td>91</td>
<td>9</td>
</tr>
<tr>
<td>American Psychological Association format for reporting statistical results</td>
<td></td>
<td></td>
</tr>
<tr>
<td>General</td>
<td>38</td>
<td>33</td>
</tr>
<tr>
<td>Elite</td>
<td>33</td>
<td>28</td>
</tr>
</tbody>
</table>

Note. Introductory general sample, n = 180 to 185; introductory elite sample, n = 54 to 55. Advanced general sample, n = 54 to 55; advanced elite sample, n = 41 to 42. Numerical values correspond to approximate number of in-class hours devoted to the topic.
For the advanced course, probability, contingency table analysis, and nonparametric procedures continued to receive relatively little emphasis. Focus on basic correlation, t tests, and one-way ANOVA understandably decreased, giving way primarily to more advanced topics in ANOVA (between-subjects factorial, within-subjects, and mixed designs; post-hoc tests and contrasts). Multiple regression also gained significant emphasis in advanced courses, but more advanced topics such as causal modeling, factor analysis, and meta-analysis continued to get minimal attention. In terms of stratification variables, the elite sample reported somewhat greater emphasis in the advanced course on simple linear regression, mixed factorial designs, and focused contrasts.

Finally, some use of computer software in statistical analysis is rapidly becoming the norm, with 87% indicating that students used a statistical package as part of one or more sta-

### Table 3. Percentages of Respondents Claiming Levels of Coverage for Statistical Techniques

<table>
<thead>
<tr>
<th>Statistical Techniques</th>
<th>Introductory Course (Coverage Level)</th>
<th>Advanced Course (Coverage Level)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0 1 2 3 Plus</td>
<td>0 1 2 3 Plus</td>
</tr>
<tr>
<td>Probability (e.g., probability &amp; counting rules, Bayes’s Theorem)</td>
<td>15 38 34 14</td>
<td>36 35 22 7</td>
</tr>
<tr>
<td>General</td>
<td>21 41 32 5</td>
<td>31 29 33 7</td>
</tr>
<tr>
<td>Elite</td>
<td>13 37 43 7</td>
<td>19 35 36 10</td>
</tr>
<tr>
<td>Analysis of contingency tables (e.g., chi square and tests of fit)</td>
<td>2 52 39 7</td>
<td>24 36 33 7</td>
</tr>
<tr>
<td>General</td>
<td>50 33 15 3</td>
<td>28 28 34 10</td>
</tr>
<tr>
<td>Elite</td>
<td>39 39 18 4</td>
<td>38 26 26 10</td>
</tr>
<tr>
<td>Other nonparametric tests (e.g., sign test, Mann–Whitney U test)</td>
<td>2 9 61 28</td>
<td>20 28 27 25</td>
</tr>
<tr>
<td>General</td>
<td>4 7 48 41</td>
<td>27 30 27 16</td>
</tr>
<tr>
<td>Elite</td>
<td>2 16 52 30</td>
<td>12 33 26 29</td>
</tr>
<tr>
<td>Correlation (e.g., Pearson r)</td>
<td>15 28 47 10</td>
<td>25 27 28 20</td>
</tr>
<tr>
<td>General</td>
<td>5 30 52 13</td>
<td>14 19 36 31</td>
</tr>
<tr>
<td>Elite</td>
<td>2 9 61 28</td>
<td>20 28 27 25</td>
</tr>
<tr>
<td>Simple, least squares regression</td>
<td>15 28 47 10</td>
<td>25 27 28 20</td>
</tr>
<tr>
<td>General</td>
<td>5 30 52 13</td>
<td>14 19 36 31</td>
</tr>
<tr>
<td>Elite</td>
<td>2 9 61 28</td>
<td>20 28 27 25</td>
</tr>
<tr>
<td>t tests on independent group (between-subjects) means</td>
<td>16 20 35 29</td>
<td></td>
</tr>
<tr>
<td>General</td>
<td>31 26 34 9</td>
<td>15 18 37 30</td>
</tr>
<tr>
<td>Elite</td>
<td>16 27 42 15</td>
<td>14 19 33 35</td>
</tr>
<tr>
<td>One-way between-subjects ANOVA</td>
<td>40 20 25 15</td>
<td>7 14 33 46</td>
</tr>
<tr>
<td>General</td>
<td>20 20 35 25</td>
<td>12 12 29 48</td>
</tr>
<tr>
<td>Elite</td>
<td>20 20 35 25</td>
<td>12 12 29 48</td>
</tr>
<tr>
<td>Factorial between-subjects ANOVA</td>
<td>40 20 25 15</td>
<td>7 14 33 46</td>
</tr>
<tr>
<td>General</td>
<td>16 12 44 28</td>
<td>16 20 35 29</td>
</tr>
<tr>
<td>Elite</td>
<td>5 13 56 25</td>
<td>14 26 36 24</td>
</tr>
<tr>
<td>One-way repeated measures analyses (including paired t tests)</td>
<td>65 22 9 4</td>
<td>16 10 33 41</td>
</tr>
<tr>
<td>General</td>
<td>65 22 9 4</td>
<td>16 10 33 41</td>
</tr>
<tr>
<td>Elite</td>
<td>56 22 9 13</td>
<td>9 2 30 58</td>
</tr>
<tr>
<td>Factorial ANOVA including one or more repeated measures</td>
<td>44 31 22 3</td>
<td>10 18 47 25</td>
</tr>
<tr>
<td>General</td>
<td>44 31 22 3</td>
<td>10 18 47 25</td>
</tr>
<tr>
<td>Elite</td>
<td>29 45 22 4</td>
<td>10 29 36 26</td>
</tr>
<tr>
<td>Post-hoc tests of differences among means (e.g., Tukey HSD)</td>
<td>77 21 3 0</td>
<td>27 20 32 21</td>
</tr>
<tr>
<td>General</td>
<td>77 21 3 0</td>
<td>27 20 32 21</td>
</tr>
<tr>
<td>Elite</td>
<td>58 25 7 9</td>
<td>17 14 33 36</td>
</tr>
<tr>
<td>Focused contrasts and comparisons</td>
<td>90 10 1 0</td>
<td>27 26 33 14</td>
</tr>
<tr>
<td>General</td>
<td>90 10 1 0</td>
<td>27 26 33 14</td>
</tr>
<tr>
<td>Elite</td>
<td>87 11 2 0</td>
<td>19 28 44 9</td>
</tr>
<tr>
<td>Multiple regression</td>
<td>74 20 5 1</td>
<td>12 14 38 35</td>
</tr>
<tr>
<td>General</td>
<td>74 20 5 1</td>
<td>12 14 38 35</td>
</tr>
<tr>
<td>Elite</td>
<td>78 13 9 0</td>
<td>7 9 35 49</td>
</tr>
<tr>
<td>Causal modeling (e.g., simple path analysis)</td>
<td>97 3 0 0</td>
<td>56 25 15 4</td>
</tr>
<tr>
<td>General</td>
<td>97 3 0 0</td>
<td>56 25 15 4</td>
</tr>
<tr>
<td>Elite</td>
<td>95 5 0 0</td>
<td>55 24 21 0</td>
</tr>
<tr>
<td>Factor analysis and/or other data reduction techniques</td>
<td>94 5 1 0</td>
<td>45 25 21 9</td>
</tr>
<tr>
<td>General</td>
<td>94 5 1 0</td>
<td>45 25 21 9</td>
</tr>
<tr>
<td>Elite</td>
<td>91 5 4 0</td>
<td>44 23 21 12</td>
</tr>
<tr>
<td>Meta-analysis</td>
<td>92 7 1 0</td>
<td>56 26 14 4</td>
</tr>
<tr>
<td>General</td>
<td>92 7 1 0</td>
<td>56 26 14 4</td>
</tr>
<tr>
<td>Elite</td>
<td>88 11 2 0</td>
<td>49 30 19 2</td>
</tr>
</tbody>
</table>

*Note.* Introductory general sample, n = 182 to 185; introductory elite sample, n = 55 to 56. Advanced general sample, n = 130 to 138; advanced elite sample, n = 42 to 43. Numerical values correspond to approximate number of in-class hours devoted to the topic.
tistical and methodology courses. Respondents listed over 25 different packages, including many innovative, instructionally oriented programs. Major professional packages and their “student versions” still dominated, however, with versions of SPSS, Inc. products mentioned by just over half of all respondents.

Discussion

The importance of collecting these data goes beyond generating an exhaustive actuarial summary of current practices. The specific content of any universal requirement deserves special scrutiny, and such information provides an important foundation for critically examining those practices (cf. Wilkinson et al., 1999). No descriptive summary directly generates prescriptions for change. Thus, our discussion raises several goal and value questions that take us beyond the data. We hope that the concerns, speculations, and suggestions we address here will serve as a catalyst for debate on these issues.

Comments on Curricular Structure

One central question involves the consequences of handling statistics and methods as separate courses rather than as an integrated sequence. In our sample, the overwhelming preference was to rely on a separate introductory statistics course, usually serving as a prerequisite to subsequent training in research methods. As noted in our results, integrated and segregated courses were associated with a number of differences in statistical coverage and emphasis. We have uncovered no empirical work, however, comparing these approaches in terms of outcomes. The heavy reliance on a segregated, introductory statistics course as a prerequisite to a methods course may simply reflect a variety of pragmatic considerations, including availability of texts, faculty expertise, and broader university staffing considerations. Nevertheless, the dominance of this approach implies some sense of its superiority that has yet to be fully explored. For example, it is quite possible that having methods training precede a statistics course could enhance students’ understanding and retention of statistical material by providing a concrete, cognitive framework to which such learning can be attached.

Despite the current absence of research on sequencing effects, there is a fair amount of work regarding different approaches for improving critical thinking and long-term statistical reasoning in introductory courses (e.g., Bradstreet, 1996; Ware & Chastain, 1991). These studies offer no specific recommendations regarding the optimal sequencing of statistics and methods within the psychology curriculum, but they do provide some indication of how certain sequences might function. It is clear from these investigations that students benefit more from their initial statistical training when taught in ways that encourage reasoning and interpretation beyond the use of statistical formulas. Whether the approach focuses on the use of writing (Beins, 1993), computers (Ware & Chastain, 1989), or word problems (Quilici & Mayer, 1996), such work could serve as a starting point for more thorough investigations into the relative merits of various sequencing or integration strategies. Given the sheer amount of student and faculty energy devoted to these core requirements, we believe that such studies—although difficult to conduct—are potentially quite valuable.

Another important question concerns the role of statistics as a prerequisite to mid-level and upper level work in the major. Our data and those of others (e.g., Messer et al., 1999) suggest a relatively low emphasis on statistical training as a prerequisite to other, nonmethods courses. Requiring statistics as a prerequisite to most upper level psychology courses would almost certainly have the impact of excluding large numbers of interested students—students whom faculty rightly believe would benefit greatly from the content of these topical offerings. However, the relatively flat, minimal prerequisite structure of most departments raises its own set of problems. Faculty often lament students’ tendency to view statistics as a kind of “foreign language requirement”—a hurdle to be passed and then forgotten. However, this is an understandable student response to a curricular structure that rarely requires statistical training for other classes.

Creative ways of reinforcing this universal requirement throughout subsequent training in the major may be critical if students are to perceive such material as anything other than an isolated requirement. For example, separate assignments within a single section of a topical course could be established for students who have already completed statistics coursework. All enrolled students might design a simple study, collecting and summarizing data for their term paper assignments. However, those students who had completed some statistical training could be required to perform more sophisticated analyses (e.g., including inferential statistics) and report them in APA style. Of course, setting what might be perceived as unequal standards within a single class presents its own set of challenges. Students may have equity concerns, and faculty may find it difficult to evaluate common projects by separate criteria. However, the general tactic of separate assignments within a single class is not without precedent. The first author, for example, has employed separate assignments within sections of an introductory psychology course without incurring negative feedback (cf. Friedrich, 1990; Friedrich & Douglass, 1998). There are certainly other approaches that merit consideration, as well. In general, we believe that new ways of reinforcing statistical training while maintaining course access to nonmajors deserve careful exploration.

Finally, our results and those of others (Messer et al., 1999; Perlman & McCann, 1999a, 1999b) indicate that a test and measurement course is rarely required, with many programs not even offering it as an elective. It seems unlikely that this absence is due simply to lack of resources, as our survey found the neglect of the measurement course to be particularly acute in the elite sample. As one reviewer suggested, it may reflect the fact that fewer new PhDs and potential instructors are receiving the necessary background as part of their own graduate training and focus (cf. Aiken et al., 1990) and may have less interest in offering such a course.

Our results further indicate that coverage of measurement issues has not been absorbed into the statistics and methods requirements. The depth in which reliability and construct validity issues are covered elsewhere in the major is beyond the scope of this survey. However, it seems unlikely that the limited coverage methods texts typically devote to these is-
sues or the mention they receive in certain elective courses touching on psychological testing are consistent with recommendations that a separate measurement course be required for all variations of the psychology major (McGovern et al., 1991). Finding new ways of infusing basic measurement instruction into the undergraduate requirement structure should be a high priority.

**Comments on Statistical Instruction**

*Statistical themes.* Taken as a whole, our results suggest that statistical instruction for undergraduate psychology majors continues to emphasize traditional approaches to analysis with relatively minimal change in response to themes and advances in the field (see Table 2). In some respects, it appears to lag behind the advances in textbooks that are available and often used in such courses. Introductory texts emphasizing effect size estimation and power analysis (e.g., Aron & Aron, 1999; Jaccard & Becker, 1997; Rosnow & Rosenthal, 1999), general linear hypothesis treatments (e.g., Aron & Aron, 1999), and even those emphasizing model comparison approaches (e.g., Lockhart, 1998) are responding to these changes in statistical practice, but instructors' self-reported emphases do not yet mirror these changes. One of the most notable deficiencies is in the coverage of confidence intervals, an area that would hardly qualify as a recent advance (cf. Loftus, 1996). Critiques of traditional null hypothesis testing and associated calls for greater emphasis on confidence interval estimation seem to be receiving the same cool reception in undergraduate instruction—at least for courses taught within psychology—that they have in professional practice.

One might wonder whether it is important to explicitly incorporate these themes into students’ earliest experiences with statistics. However, to the extent that the introductory course establishes the conceptual framework around which students will organize their statistical knowledge, it is tempting to make the case that early coverage is especially critical and that it might appropriately displace some of the instruction time normally spent on specific statistical tests. Although highlighting these themes in data analysis has often been discussed in terms of top-down solutions linked to advances in the field (see Sherman et al., 1996). Our own informal review suggests that the artificial distinction between ANOVA and regression techniques tends to be reinforced in most undergraduate methods and statistics texts. We suspect that students can (and often do) walk away from such treatments mistakenly thinking that experimental versus correlational study designs are determined by the choice of analytical technique and that “test” reliability and validity are primarily a concern of the latter. Indeed, some text authors attempt to warn readers of this potential confusion (e.g., Goodwin, 1998). We believe that a greater emphasis on general linear model approaches has the potential both to help students access a broader range of empirical reports and independent projects and to pave the way for a better integration of measurement issues into the curriculum.

A *Not-So-Modest Proposal*

Realistically, only so much can be asked of students and instructors in introductory statistics courses. Psychology majors often enter such courses with great apprehension and struggle to master a way of thinking that may represent an abrupt departure from their previous analytical and mathematical training. Our data indicate that instructors view many topics important for reading access to even the less complex journal literature, as well as techniques likely to be incorporated into senior theses and research projects, to be the proper domain of an advanced statistics course. Indeed, even the previously noted thematic controversies and advances are topics instructors said they are (or would be) inclined to address primarily in a more advanced course. Although we have argued that instructors can and should address some of these issues from the very beginning of statistical instruction, these facts raise the question of whether training beyond the introductory course should become the norm rather than the exception for psychology majors.

The question of whether to increase major requirements in methodology and statistics is a difficult one. Only a small mi-
Minority of undergraduate psychology majors pursue graduate training or other occupations for which more advanced undergraduate training might be particularly useful. Moreover, for the liberal arts context within which most psychology majors pursue their degrees, narrowly technical preprofessional training is philosophically problematic. Yet with quantitative and methodological skills being among the particular strengths that bachelor’s-level majors might bring to their careers in applied settings (Clough, 1993), it may be unwise to treat more extensive statistical coverage as a concern only for graduate school-bound students. The low availability of advanced quantitative offerings—either as requirements or as electives—reported in our survey has important consequences for the large number of undergraduates who might benefit from developing stronger research skills and for the much smaller number preparing for graduate education.

Taken as a whole, we view these survey results as making a strong case for offering a dedicated, advanced elective course beyond the typical 1-year sequence. Such a course could even be conceived of as a potential requirement in certain programs. In effect, the course could be taken in lieu of the widely recommended course in tests and measurement. Given that our data suggest a measurement course is rarely required, an advanced course of the sort we describe subsequently would not expand methodology requirements beyond the three-course program (i.e., statistics, methods, and measurement) already strongly endorsed in curricular guidelines (Brewer et al., 1993; McGovern et al., 1991). We propose that such an advanced, hybrid course might focus on several goals:

1. A balanced exposure to basic ANOVA and regression techniques as part of a model comparison—general linear model approach, providing a more solid foundation for topics in measurement and nonexperimental research than is typical of advanced courses.
2. Explicit coverage of measurement issues in reliability and construct validation with links to broader design and analysis concerns, but with surveys of standardized tests left to more topically focused courses.
3. A strong emphasis on basic thematic issues such as confidence intervals, effect size, power analysis, and exploratory displays that cut across particular tests and designs.
4. A conceptual treatment of topics such as basic causal modeling and meta-analysis (e.g., Judd, Smith, & Kidder, 1991) to better prepare students for their encounters with current literature reviews and the increasing diversity of techniques in primary reports (cf. Sherman et al., 1999).

Although we suspect that such a proposal might be seen as targeting a graduate school-bound population, we wish to emphasize that this is not the case. Indeed, what we propose emphasizes a conceptual rather than computational focus that is a far cry from the typical introductory graduate sequence. An undergraduate course should provide a good foundation that facilitates and organizes learning in potential graduate coursework. However, we believe such a class should prioritize the conceptual and analytical skills that can equally enhance undergraduate students’ experiences in their elective coursework, their placements in community agencies doing evaluation work, their research activities, and their eventual careers.

We can only speculate regarding the reasons for the patterns of instruction noted in our survey—patterns we have argued might require significant change in light of recent guidelines for statistical practice in psychology (Wilkinson et al., 1999). It seems likely that the willingness of textbook authors and publishers to embrace new approaches plays a significant role (Dillon, 1999). An even greater contributor may be the inertia of instructors’ own past training in statistics. No doubt like most readers of this report, our primary graduate specializations and professional work are in content areas other than statistics and measurement, and a major challenge to reconsidering one’s approach to statistical instruction involves reexamining one’s training and habits as a professional (cf. Zuckerman, Hodgins, Zuckerman, & Rosenthal, 1993). For example, we suspect that the relatively low coverage or integration of regression techniques with ANOVA topics reflects the greater emphasis on the latter in the training of experimental psychologists who often teach the statistics and design courses. For busy faculty, the priority of keeping abreast of broader developments in statistical practice may be understandably much lower than for staying current in their topical specialty areas.

An acquaintance, in discussing his struggle to keep up with the literature in the various areas in which he taught, once remarked that the nicest thing about teaching statistics courses is that the content never changes. Yet work with direct implications for undergraduate statistical instruction continually emerges, making even our present recommendations perhaps seem overly cautious (e.g., Bear, 1995; Frick, 1998; Wilcox, 1998). Our survey results indicate that undergraduate statistical instruction is indeed heavily tied to tradition, but they also suggest ways psychology departments might seek to update and invigorate this core area of the major.

Space does not permit a thorough discussion of how instructors could be encouraged to make such changes, and it does not permit consideration of changes in teaching strategies that might be needed to effectively deliver a new curriculum such as the one we have proposed. Nevertheless, we share the views of the APA Task Force on Statistical Inference (Wilkinson et al., 1999) in calling for the incorporation of such changes into standard practice, and we hope that a commitment to this goal will lead to innovations in curriculum and strategies for teaching that are accompanied by careful empirical evaluation. Collectively, teachers of undergraduates have a golden opportunity to foster this “revolution” by changing statistical practice from the bottom up—an opportunity we believe can be exploited to the benefit of both the discipline and its students.

References


Clough, J. (1993). Which knowledge and skills should psychology graduates have? Balancing the needs of the individual, employers, the science and the profession. *Australian Psychologist*, 28, 42–44.


Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, 5, 161–171.


**Notes**

1. Evelyn Buday is now at Alleghany College, Meadville, PA.
2. This research was supported by a faculty Atkinson Research Grant from Willamette University.
3. Portions of this article were presented at the American Psychological Association Convention, San Francisco, August 1998.
4. Send correspondence to James Friedrich, Department of Psychology, Willamette University, 900 State Street, Salem, OR 97301; e-mail: jfriedri@willamette.edu.
Core Terms in Undergraduate Statistics

R. Eric Landrum
Boise State University

I analyzed 3 introductory statistics textbooks to generate a listing of key terms and concepts. After removing duplications, 374 unique terms formed the master list. A national sample of introductory statistics instructors (N = 190) evaluated portions of the master list, rating each item on a scale ranging from 1 (not at all important) to 3 (extremely important). I list the Top 100 items and offer recommendations about the usefulness of the importance listing in facilitating statistics instruction.

An undergraduate course in statistics continues to be an essential component in the curriculum for the psychology major. After reviewing curricula from hundreds of colleges and universities nationwide, Perlman and McCann (1998) concluded that the core of the undergraduate major comprises introductory psychology, statistics, and a capstone course. Instructors of statistics face a host of challenges (Conners, Mccown, & Roskos-Ewoldsen, 1998; Hastings, 1982). For instance, Conners et al. listed four unique challenges in teaching the statistics course: (a) motivating students, (b) overcoming math anxiety, (c) avoiding performance extremes, and (d) making learning last. Given the prevalence of the statistics course and the unique challenges of teaching it, any resource available to facilitate instructor performance in teaching statistics would be highly valuable.

This study shares similarities with previous studies (e.g., Giesbrecht, Sell, Scialfa, Sandsals, & Ehlers, 1997). Those authors had 18 professors from four different disciplines rate the importance of topics in research methods and statistics. Giesbrecht et al. obtained topic listings from previous research, review of statistics and research methods texts, and course descriptions, and “found interdepartmental agreement on the relative importance of 97% of the statistics topics, with 74% of those topics deemed to be important” (p. 245). In this study, I focused on statistics instruction at the undergraduate level only and within psychology only, and I base my conclusions on the responses of a national sample of undergraduate statistics instructors.

My goal was to generate a master list of statistical terms relevant to an undergraduate statistics course and then to ask statistics instructors to rate the importance of those terms in that context. Other researchers have successfully used this approach with introductory psychology terms (Boneau, 1990; Griggs & Mitchell, 2002; Landrum, 1993; Quereshi, 1993; Quereshi & Sackett, 1977; Zechmeister & Zechmeister, 2000). For instance, Zechmeister and Zechmeister performed a content analysis of terms using the glossaries of 10 introductory psychology textbooks and generated a list of 2,505 unique terms. After dividing the list into smaller sections, Zechmeister and Zechmeister then sent the sections to a sample of introductory psychology instructors for importance ratings. These top terms are construed as the core terms in introductory psychology.

There are potential benefits in having a set of core terms for an undergraduate statistics course. New instructors of statistics would have a valuable reference in helping determine course content. All statistics instructors would have the ability to compare their decisions about course content (especially across multiple sections) to the importance ratings of a national sample of statistics instructors. Regarding textbook selection, other approaches have been helpful in choosing a statistics book, such as assessing readability and writing style (Harwell, Herrick, Curtis, Mundfrom, & Gold, 1996). A core terms listing for statistics might also assist instructors in the textbook selection process. Instructors face time constraints regarding the content and topics of the course versus the length of the semester. Importance ratings might help instructors (or even statistics textbook authors) determine those topics that are less important. Making this type of informed choice might lead an instructor to come back to a topic at the end of a semester, time permitting. Thus, there are multiple plausible advantages to having a list of core terms in statistics rated by importance.

Method

Participants

Psychology department chairs received a cover letter explaining the purpose of this study, a survey, and a self-addressed business-reply envelope. I asked the department chair to forward the survey to an undergraduate statistics instructor. I mailed 814 surveys to colleges and universities in the United States that offered an undergraduate degree in psychology—I obtained this contact information from the American Psychological Association Office of Research. A total of 190 instructors responded with usable data, yielding a response rate of 23.3%.

Materials and Procedure

Prior to surveying statistics instructors for importance ratings, I developed a master list of terms; I describe that process here and the materials used to survey statistics instructors.

Generation of the master list. I selected a convenience sample of statistics texts (i.e., those that were on my bookshelf): Gravetter and Wallnau (1999), Pagano (1998), and Spatz (1997). The important details for each text follow: Gravetter and Wallnau: 458 pages, 16 chapters, 210 key terms; Pagano: 548 pages, 19 chapters, 238 key terms; Spatz: 488 pages, 14 chapters, 167 key terms. Because not all three books contained glossaries, I could not use previous content analysis techniques (e.g., Zechmeister & Zechmeister, 2000). Instead, I performed a page-by-page content analysis of all three books and selected terms in bold face type or italics. In some instances I used the heading of a particular section. I previously used a similar technique in my examination of introductory psychology textbooks (Landrum, 1993). At times, I had to make decisions about similar terms. For example, “line of best fit” and “best-fitting line” mean essentially the same thing; I eliminated one from the final list of terms. This decision was subjective; if terms were highly
similar but not identical, I erred on the side of caution and included both terms.

The three statistics texts yielded 615 terms; 374 unique terms emerged to constitute the list of statistics terms. After some consideration, I decided not to compare terms chapter by chapter, as previously done in the content analysis of introductory psychology textbooks (e.g., Landrum, 1993; Zechmeister & Zechmeister, 2000). For instance, one book had a separate chapter for measures of central tendency and variability, whereas another combined the two topics into one chapter. Of the 615 terms coded, 44 terms appeared in all three books.

Importance ratings by statistics instructors. I believed that 374 statistical terms was too many to ask any one person to evaluate, so I randomly divided the listing of terms into three separate, mutually exclusive forms. Thus, each instructor rated a list of 124 or 125 terms, presented alphabetically. On average, 63.3 instructors (SD = 13.0) rated each list of terms. Partic-ipants rated each concept using a 4-point importance scale from 0 (not at all important), 1 (slightly important), 2 (moderately important), to 3 (extremely important). I instructed participants to leave an item blank if it was unclear, and I calculated a mean rating for each term. The term normal curve was the only term to receive a unanimous rating of 3.0 (for the means and standard deviations of the top 100 terms, see Table 1).

Table 1. Means and Standard Deviations for the Top 100 Terms

<table>
<thead>
<tr>
<th>Item</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal curve</td>
<td>3.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Statistically significant</td>
<td>2.98</td>
<td>0.13</td>
</tr>
<tr>
<td>Bell-shaped curve</td>
<td>2.96</td>
<td>0.18</td>
</tr>
<tr>
<td>Significance level</td>
<td>2.96</td>
<td>0.18</td>
</tr>
<tr>
<td>Hypothesis testing</td>
<td>2.96</td>
<td>0.19</td>
</tr>
<tr>
<td>Normal distribution</td>
<td>2.96</td>
<td>0.19</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>2.96</td>
<td>0.19</td>
</tr>
<tr>
<td>Sample</td>
<td>2.94</td>
<td>0.22</td>
</tr>
<tr>
<td>Alpha level</td>
<td>2.94</td>
<td>0.23</td>
</tr>
<tr>
<td>Mean</td>
<td>2.93</td>
<td>0.26</td>
</tr>
<tr>
<td>Null hypothesis</td>
<td>2.93</td>
<td>0.37</td>
</tr>
<tr>
<td>Central tendency</td>
<td>2.92</td>
<td>0.26</td>
</tr>
<tr>
<td>Inferential statistics</td>
<td>2.92</td>
<td>0.31</td>
</tr>
<tr>
<td>Variability</td>
<td>2.92</td>
<td>0.31</td>
</tr>
<tr>
<td>Arithmetic mean</td>
<td>2.91</td>
<td>0.28</td>
</tr>
<tr>
<td>Correlation</td>
<td>2.90</td>
<td>0.29</td>
</tr>
<tr>
<td>Pearson correlation</td>
<td>2.90</td>
<td>0.29</td>
</tr>
<tr>
<td>Dependent variable</td>
<td>2.89</td>
<td>0.34</td>
</tr>
<tr>
<td>Two-tailed probability</td>
<td>2.89</td>
<td>0.44</td>
</tr>
<tr>
<td>Positive correlation</td>
<td>2.88</td>
<td>0.37</td>
</tr>
<tr>
<td>Data</td>
<td>2.88</td>
<td>0.38</td>
</tr>
<tr>
<td>Hypothesis</td>
<td>2.88</td>
<td>0.45</td>
</tr>
<tr>
<td>t test</td>
<td>2.87</td>
<td>0.51</td>
</tr>
<tr>
<td>Descriptive statistics</td>
<td>2.86</td>
<td>0.34</td>
</tr>
<tr>
<td>Variance</td>
<td>2.86</td>
<td>0.34</td>
</tr>
<tr>
<td>Negative correlation</td>
<td>2.86</td>
<td>0.43</td>
</tr>
<tr>
<td>Not significant</td>
<td>2.85</td>
<td>0.44</td>
</tr>
<tr>
<td>Variable</td>
<td>2.84</td>
<td>0.42</td>
</tr>
<tr>
<td>Population</td>
<td>2.84</td>
<td>0.52</td>
</tr>
<tr>
<td>Statistic</td>
<td>2.84</td>
<td>0.52</td>
</tr>
<tr>
<td>Level of significance</td>
<td>2.83</td>
<td>0.37</td>
</tr>
<tr>
<td>Critical values</td>
<td>2.83</td>
<td>0.46</td>
</tr>
<tr>
<td>Type I error</td>
<td>2.82</td>
<td>0.46</td>
</tr>
<tr>
<td>Degrees of freedom</td>
<td>2.81</td>
<td>0.43</td>
</tr>
</tbody>
</table>

Table 1 (Continued)

<table>
<thead>
<tr>
<th>Item</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
<td>2.81</td>
<td>0.44</td>
</tr>
<tr>
<td>Significant effect</td>
<td>2.81</td>
<td>0.53</td>
</tr>
<tr>
<td>Rejection region</td>
<td>2.81</td>
<td>0.57</td>
</tr>
<tr>
<td>t test for independent-samples design</td>
<td>2.81</td>
<td>0.57</td>
</tr>
<tr>
<td>One-way ANOVA</td>
<td>2.80</td>
<td>0.44</td>
</tr>
<tr>
<td>Statistical inference</td>
<td>2.79</td>
<td>0.46</td>
</tr>
<tr>
<td>Two-tailed test of significance</td>
<td>2.79</td>
<td>0.51</td>
</tr>
<tr>
<td>t test for independent groups</td>
<td>2.79</td>
<td>0.58</td>
</tr>
<tr>
<td>t statistic</td>
<td>2.79</td>
<td>0.63</td>
</tr>
<tr>
<td>Standard error of the mean</td>
<td>2.77</td>
<td>0.55</td>
</tr>
<tr>
<td>Critical region</td>
<td>2.76</td>
<td>0.50</td>
</tr>
<tr>
<td>Standard error</td>
<td>2.75</td>
<td>0.58</td>
</tr>
<tr>
<td>ANOVA</td>
<td>2.75</td>
<td>0.63</td>
</tr>
<tr>
<td>Inferential process</td>
<td>2.74</td>
<td>0.57</td>
</tr>
<tr>
<td>Alternative hypothesis</td>
<td>2.74</td>
<td>0.60</td>
</tr>
<tr>
<td>F ratio</td>
<td>2.73</td>
<td>0.48</td>
</tr>
<tr>
<td>Deviation</td>
<td>2.73</td>
<td>0.58</td>
</tr>
<tr>
<td>Distribution of sample means</td>
<td>2.73</td>
<td>0.59</td>
</tr>
<tr>
<td>Student's t test</td>
<td>2.73</td>
<td>0.66</td>
</tr>
<tr>
<td>Linear relationship</td>
<td>2.72</td>
<td>0.55</td>
</tr>
<tr>
<td>Independent-samples design</td>
<td>2.72</td>
<td>0.64</td>
</tr>
<tr>
<td>z score transformation</td>
<td>2.72</td>
<td>0.67</td>
</tr>
<tr>
<td>Random</td>
<td>2.71</td>
<td>0.61</td>
</tr>
<tr>
<td>Random assignment</td>
<td>2.71</td>
<td>0.62</td>
</tr>
<tr>
<td>Sampling error</td>
<td>2.70</td>
<td>0.51</td>
</tr>
<tr>
<td>Correlational method</td>
<td>2.70</td>
<td>0.60</td>
</tr>
<tr>
<td>z score</td>
<td>2.70</td>
<td>0.72</td>
</tr>
<tr>
<td>Null-hypothesis population</td>
<td>2.68</td>
<td>0.65</td>
</tr>
<tr>
<td>Frequency</td>
<td>2.67</td>
<td>0.54</td>
</tr>
<tr>
<td>Independent groups design</td>
<td>2.67</td>
<td>0.58</td>
</tr>
<tr>
<td>Frequency distribution</td>
<td>2.67</td>
<td>0.70</td>
</tr>
<tr>
<td>Independent variable</td>
<td>2.66</td>
<td>0.63</td>
</tr>
<tr>
<td>Type II error</td>
<td>2.64</td>
<td>0.64</td>
</tr>
<tr>
<td>One-tailed probability</td>
<td>2.62</td>
<td>0.67</td>
</tr>
<tr>
<td>Random selection</td>
<td>2.62</td>
<td>0.72</td>
</tr>
<tr>
<td>Nondirectional hypothesis</td>
<td>2.61</td>
<td>0.64</td>
</tr>
<tr>
<td>Sampling distribution</td>
<td>2.61</td>
<td>0.71</td>
</tr>
<tr>
<td>Estimated population standard deviation</td>
<td>2.60</td>
<td>0.63</td>
</tr>
<tr>
<td>Overall mean</td>
<td>2.60</td>
<td>0.67</td>
</tr>
<tr>
<td>Correct decision</td>
<td>2.60</td>
<td>0.77</td>
</tr>
<tr>
<td>Sampling distribution of the mean</td>
<td>2.59</td>
<td>0.72</td>
</tr>
<tr>
<td>Sampling distributions of a statistic</td>
<td>2.59</td>
<td>0.72</td>
</tr>
<tr>
<td>Regression</td>
<td>2.59</td>
<td>0.81</td>
</tr>
<tr>
<td>Causation</td>
<td>2.58</td>
<td>0.56</td>
</tr>
<tr>
<td>Scatterplot</td>
<td>2.58</td>
<td>0.61</td>
</tr>
<tr>
<td>Sum of squares</td>
<td>2.57</td>
<td>0.63</td>
</tr>
<tr>
<td>Positive relationship</td>
<td>2.55</td>
<td>0.67</td>
</tr>
<tr>
<td>Sampling distribution of f</td>
<td>2.55</td>
<td>0.73</td>
</tr>
<tr>
<td>Sum of squared deviations</td>
<td>2.55</td>
<td>0.77</td>
</tr>
<tr>
<td>Test statistic</td>
<td>2.55</td>
<td>0.82</td>
</tr>
<tr>
<td>Chi-square distribution</td>
<td>2.54</td>
<td>0.65</td>
</tr>
<tr>
<td>Between-groups sum of squares</td>
<td>2.54</td>
<td>0.75</td>
</tr>
<tr>
<td>Simple random sample</td>
<td>2.54</td>
<td>0.75</td>
</tr>
<tr>
<td>Population variance</td>
<td>2.53</td>
<td>0.72</td>
</tr>
<tr>
<td>Random sampling</td>
<td>2.52</td>
<td>0.78</td>
</tr>
<tr>
<td>t distribution</td>
<td>2.51</td>
<td>0.69</td>
</tr>
<tr>
<td>Chi-square statistic</td>
<td>2.51</td>
<td>0.78</td>
</tr>
<tr>
<td>One-tailed test of significance</td>
<td>2.51</td>
<td>0.78</td>
</tr>
<tr>
<td>Probability</td>
<td>2.50</td>
<td>0.70</td>
</tr>
<tr>
<td>Standard score</td>
<td>2.50</td>
<td>0.75</td>
</tr>
<tr>
<td>F distribution</td>
<td>2.50</td>
<td>0.83</td>
</tr>
<tr>
<td>Distribution of scores</td>
<td>2.49</td>
<td>0.64</td>
</tr>
<tr>
<td>ANOVA summary table</td>
<td>2.49</td>
<td>0.74</td>
</tr>
<tr>
<td>Treatment</td>
<td>2.49</td>
<td>0.75</td>
</tr>
<tr>
<td>Levels/treatments</td>
<td>2.49</td>
<td>0.78</td>
</tr>
<tr>
<td>Subjects/participants</td>
<td>2.49</td>
<td>0.81</td>
</tr>
</tbody>
</table>
Discussion

The results of this study, as presented in Table 1, provide an interesting look at what statistic instructors consider important. This Top 100 list may be construed as a first approximation of the core of introductory statistics in psychology. These results may also help instructors decide what the important terms to cover in class are and what terms are secondary. The approach of this study is unique because only undergraduate statistics instructors rated the importance of terms. Departments that offer multiple sections of statistics taught by different instructors using different textbooks might find these results useful in an effort to coordinate consistent coverage of the various content areas of statistics. The results of this study are more specific to psychology instructors than those from a previous study (Giesbrecht et al., 1997).

For some semesters, instructors may not cover certain chapters of information they consider less important; other semesters they may cover those chapters. What topics sometimes get deleted due to time constraints? The listing presented in Table 1 should help instructors and authors to answer these questions by understanding the core terms in the context of relative importance.

This study has limitations with regard to the methodological approach used, including the convenience sample of textbooks. In addition, in the instances where terms were highly similar but not identical, I included all terms. More work in this area is warranted to continue to verify the core terms. The listing provided here can be valuable for instructors in designing and making choices about the coverage of topics in an introductory course in statistics.

References


Promoting Conceptual Understanding of Statistics: Definitional Versus Computational Formulas

Katarina Guttmannova, Alan L. Shields, and John C. Caruso
University of Montana

Computer applications have replaced hand calculations as the relevant procedural skill for most of the statistical techniques in introductory statistics courses. Therefore, definitional formulas should replace computational formulas because only the former contribute to conceptual understanding. A review of 12 introductory statistics textbooks indicated that they emphasized computational formulas, particularly for complex techniques and exercises. We argue that the presentation and use of computational formulas is counterproductive to the learning goals of statistics courses and provide recommendations for instructors to facilitate the use of definitional formulas.

In the precomputer era, when students performed calculations by hand, computational formulas provided heuristics for difficult computations but sacrificed understanding of the underlying statistical concepts. Currently, however, 69% of psychology departments use computers in introductory statistics courses, and 90% use them for data analysis at some point in their curriculum (Bartz & Sabolik, 2001). Computers perform most statistical analyses. Thus we question the continued need for computational formulas and advocate instead for definitional formulas computed using handheld calculators.

Although hand calculations are less tedious when using computational formulas, definitional formulas provide students with some perspective as to what the statistic is, when to use it, and how to interpret it. Consider, for example, the definitional,

\[ s^2 = \frac{\sum (X_i - M)^2}{N}, \]

and computational,
Statistics and Methodology Courses: Interdepartmental Variability in Undergraduate Majors' First Enrollments

Joan B. Lauer, D. W. Rajecki and Karl A. Minke

Teaching of Psychology 2006 33: 24
DOI: 10.1207/s15328023top3301_6

The online version of this article can be found at:
http://top.sagepub.com/content/33/1/24

Published by:
$SAGE$
http://www.sagepublications.com

On behalf of:

Society for the Teaching of Psychology

Additional services and information for Teaching of Psychology can be found at:

Email Alerts: http://top.sagepub.com/cgi/alerts
Subscriptions: http://top.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Jan 1, 2006

What is This?
Statistics and Methodology Courses: Interdepartmental Variability in Undergraduate Majors’ First Enrollments

Joan B. Lauer and D. W. Rajecki
Indiana University–Purdue University Indianapolis

Karl A. Minke
University of Hawaii at Manoa

Transcripts of 784 psychology alumni from 4 universities revealed that students’ timing of first enrollments in a statistics or methodology course was the result of an interaction between personal preferences and differences in program requirements. Where only a single methodological course was mandatory, first enrollments were especially late in students’ careers. In departments having multiple methodological requirements, negative correlations between the timing of first method course versus first content course (e.g., clinical, developmental) indicated that the precedence of 1 category meant a delay for the other. We offer recommendations for increasing students’ direct early exposure to scientific methodology in psychology.

National surveys of psychology faculty and reviews of course listings indicate interdepartmental variability in undergraduate research opportunities (Perlman & McCann, 1999b, 2005). Our project extends this literature by profiling the programs of selected departments circa 1995 and connecting differences in prerequisites and requirements with the actual timing of students’ first enrollments in methodological and other psychology courses. Interdepartmental variation is worth extended study because it bears on the issue of differing faculty and student evaluations of particular course content.

From a faculty perspective, educators in psychology encourage early direct involvement in statistics and methodology so that resultant skills can enhance other spheres of study (cf. Berthold, Hakala, & Goff, 2003; Perlman & McCann, 2005). According to this view, courses with a statistical or methodological focus promote numeracy, technical literacy, and problem-solving abilities (Graham, 1998; Hayes, 1997; Kruger & Zechmeister, 2001; Levy, Burton, Mickler, & Vigorito, 1999). Furthermore, for students who do not go on to achieve professional status in the discipline, such proficiencies convey an advantage in the new-economy job market (Knotts, 2002; Lorig, 1996; Murray, 1997; Volpe, 2002).

For our purpose it is useful to note the long tradition of faculty commitment to teach statistics and methodology. By the late 1940s, a majority of U.S. universities listed undergraduate psychology courses with titles such as experimental, statistics, or tests and measurement, and a near-majority offered research (Sanford & Fleishman, 1950). Subsequent national conferences and reports continued this emphasis. Beginning with the 1951 Cornell Conference, there followed the 1960 Michigan Conference, the 1988 Association of American Colleges Report, the 1991 St. Mary’s Conference, and, most recently, the Task Force on Undergraduate Psychology Major Competencies. All these advisory bodies highlighted statistics and methodology in versions of a model curriculum (American Psychological Association Task Force on Undergraduate Psychology Major Competencies, 2002; Brewer, 1997).
Even so, from a student perspective, there is much more interest in the area of personality and social than quantitative and experimental. One observer remarked on psychology instructors’ “difficulties of dealing with undergraduates whose palms sweat at the thought of numbers” (Beins, 1992, p. 526); another author went so far as to coin the term “statisticophobia” (Dillon, 1982). Recent research supports these observations. For example, a survey in our department found that, regardless of career plans, freshmen psychology majors already showed a preference for human interest courses (e.g., abnormal, developmental) over methodological courses (e.g., statistics, laboratory; Rajekcik, Appleby, Williams, Johnson, & Jeschke, 2005). Concerning alumni, Johanson and Fried (2002) found that their former students judged statistics and laboratory courses less valuable compared to abnormal and developmental titles. Similarly, a national survey of 1992 psychology baccalaureates found that a majority rated research and statistical experiences as “not helpful” in obtaining their primary job. Alternatively, those graduates gave clinical, developmental, and social courses higher endorsements (Grocer & Kohout, 1997).

As is the case with the tradition of professional commitment to teach statistics and methodology, a strain of student resistance to the scientific method is traceable to the middle of the last century. In 1951, undergraduates assigned preference ranks to 10 topics from a general psychology course. The topics of personality or adjustment and social psychology ranked 1 and 3, respectively, whereas experimental and mental tests and measurement ranked 7 and 9. All topics were ranked again by students in 1968, revealing a strong association (r = .92) across relative preferences (Anderson, 1970).

Subsequently, Adelman (1995) reported separate, decade-long national surveys—the first running from 1972 to 1984, the other from 1981 to 1993—to determine the percentages of all baccalaureates who had completed various psychology courses beyond the introductory level. By our calculation, developmental enjoyed the highest rate of exposure (unweighted M = 22.7%) followed by social (15.0%) and abnormal (13.6%). Far fewer undergraduates took experimental (4.7%), quantitative (3.6%), or psychometrics (1.5%). There was a strong association between levels of course popularity across the two surveys (r = .93).

Given these trends, professors do not rely solely on students’ intrinsic motivation in the matter of teaching psychology’s breadth and strengths. The words numeracy and technical literacy may be an advisor’s truthful promises, but the words requirement and prerequisite are a curriculum committee member’s useful prods. The presence of departmental requirements notwithstanding, many undergraduate psychology majors can somewhat shape their curricular experiences in response to personal priorities. We employed course mapping (or transcript analysis) to empirically identify enrollment sequences, where the objective was the actual chronology of students’ exposure to various titles. The promise of this general approach is a precise record of what was studied, where, when, and by whom (Adelman, 1995).

In hypothetical (if–then) terms, if in some program a student has the opportunity to select early and frequent enrollments in popular content courses, then these classes necessarily displace less appealing statistics and methodology offerings to later segments of her or his career. The resulting delay of statistical and methodological courses is out of line with faculty expectations of early direct exposure to scientific methods. Therefore, we document how differences in psychology programs relate to varieties of students’ actual course patterns or sequences.

Method

Because our project concerned departmental variability in educational outcomes, it was important to document similarities and differences among settings, programs, and participants. This section opens with sketches of the institutions, course listings, and students in question. Next, we provide details of our course mapping procedures.

Universities

Convenience samples of psychology graduates’ transcripts came from four universities—two urban and two traditional. One urban university program was the BA psychology major at the University of Hawaii Manoa (UHM). The second urban program was the BA/BS psychology major at Indiana University–Purdue University Indianapolis (IUPUI). During the study period UHM and IUPUI were, respectively, second- and third-tier national universities (“Best National Universities,” 1997). In terms of traditional universities, the third program was the BA/BS psychology major at Indiana University Bloomington (IUB), and the fourth was the BA psychology major at the University of Colorado Boulder (UCB). Both traditional institutions were second-tier national universities (“Best National Universities,” 1997).

Undergraduate Psychology Programs

UHM: University of Hawaii at Manoa 1995–1997 General and Graduate Information Catalog. Following a single introductory course, students had to complete four courses from a list of seven topics: experimental, psychobiology, developmental, social, personality and transpersonal, adjustment and clinical, and applied. Requirements also included one lower division quantitative course. Students could choose from one of two methodology titles or statistics. There were no prerequisites for the methodology courses; statistics required 2 years of high school algebra.

IUPUI: Purdue University School of Science Bulletin 1993–1995. Following a two-course introductory sequence, students had to complete at least six courses from a list that included developmental, physiological, cognition, learning, motivation, social, abnormal, and personality. Requirements also included two lower division quantitative courses, a laboratory and statistics. There were no prerequisites for the laboratory course. Regarding the statistics course, BA majors had to have completed 3 credits of college mathematics that included algebra or trigonometry; BS majors needed 3 credits of college mathematics beyond algebra and trigonometry. Finally, an upper division laboratory represented a third methodological requirement.
ILIB: Indiana University Bulletin—College of Arts and Sciences 94/96. Introductory psychology involved a two-course sequence for regular students or an intensive one-course option for honors students. All students needed to complete at least two courses from a list that included childhood and adolescence, social, and abnormal, and two from a menu including learning, behavioral neuroscience, and cognitive. Requirements also included two lower division quantitative courses, experimental and statistics. The experimental course required college basic algebra; the statistics course required finite mathematics or a survey of calculus. Finally, BA students had to complete a third methodology course in the form of an upper division laboratory, and BS students faced two upper division methodology courses.

UCB: Psychology requirements worksheet (for students declaring the major between July 4, 1994 and May 15, 1998). Three lower division courses were mandatory: general, biological, and statistics and methods. The statistics and methods course had no prerequisites. Several lower division substantive courses were available for credit, including adjustment, child and adolescent, and social problems. Requirements also included upper division content credits from the categories of social and clinical and biopsychology. Finally, students had to complete a second methodology course, chosen from a list that included specialties in learning, cognitive, and social.

Summary. The major course requirements within the four programs had many things in common. For example, all presented undergraduates with opportunities for exposure to a wide range of subdisciplines in the field of psychology and, to a considerable extent, all students could exercise choices in enrollments. However, there were also noteworthy differences. First, the UHM program demanded the completion of only one quantitative course, whereas all other programs required two or more quantitative courses. Second, UCB was the only program that required a specific lower division biological psychology course. All other programs offered options. Third, two programs (UHM and UCB) had a single introductory course, whereas two (IUPUI and IUB) had a two-course sequence. Fourth, two programs (IUPUI and IUB) demanded college-level mathematics as prerequisites for certain quantitative courses, whereas two programs (UHM and UCB) demanded none. We expected the course maps of alumni from those institutions to reflect such differences.

Student Characteristics

We restricted samples to so-called traditional psychology majors (Metzner, Rajeczi, & Lauer, 1994). Traditional majors are alumni who entered a 4-year institution, declared the psychology major there, and eventually graduated without earning transfer credit. Differences in sample sizes are incidental and are due in part to the sheer size of programs. Furthermore, UHM and IUPUI served many majors with transfer credits who, by definition, were not traditional majors.

UHM majors. The 75 UHM graduates (40% men) in our sample ranged in age from 21.3 to 36.0 years, with a mean of 23.7. Ethnic categories were 35% Japanese; 13% Hawaiian; 12% mixed; 11% White; 8% each Chinese, Filipino, and Korean; and 5% unspecified. Inclusive psychology GPA point average (GPA; any and all courses with a letter grade from A to F) at graduation ranged from 2.50 to 4.00, with a mean of 3.21. These students graduated between December 1995 and May 1997.

IUPUI majors. The 104 IUPUI major graduates (53 BA, 51 BS; 25% men) ranged in age from 21.6 to 55.5 years, with a mean of 28.5. Ethnic categories were 86% White, 11% Black, 2% Asian, and 1% other. Inclusive psychology GPA at graduation ranged from 1.95 to 4.00, with a mean of 3.16. These students graduated between December 1995 and May 1999.

IUB majors. The 354 IUB graduates (286 BA, 68 BS; 35% men) ranged in age from 20.8 to 31.3 years, with a mean of 22.7. Ethnic categories were 95% White; 3% Black; and the remainder Hispanic, Asian American, Native American, or unspecified. Inclusive psychology GPA at graduation ranged from 2.10 to 4.00 with a mean of 3.29. These students graduated between May 1995 and December 1997.

UCB majors. The 251 UCB graduates (32% men) ranged in age from 20.9 to 27.8 years, with a mean of 22.8. Ethnic categories were 82% White, 7% Asian American, 5% Hispanic, 4% unspecified, and 2% Black. Inclusive psychology GPA at graduation ranged from 2.18 to 4.00 with a mean of 3.28. These students graduated between May 1995 and December 1997.

Transcripts

At IUPUI, the first (JBL) and second (DWR) authors served as departmental counselors whose official duties included routine audits of all psychology majors’ transcripts. In conjunction with other local lines of research, from 1995 to 1999 we obtained copies of the records of successful candidates for graduation. We also received institutional review board (IRB) approval for the use of IUPUI transcript information for research, including a waiver of students’ informed consent. Subsequently, opportunities arose for obtaining additional sets of records from UHM, IUB, and UCB. To make the four samples comparable, our pilot requests to those chairs or registrars specified graduates with no transfer credits and graduation dates from 1995 to 1997, which yielded sufficient useful cases. The documents we received did not contain identifiers, and we obtained extended IRB approval to use the three outside transcript collections.

For information about departments, we also obtained relevant course catalogs, academic bulletins, and departmental documents from the three schools outside IUPUI. This approach capitalized on certain existing departmental conditions at a given point in time. Certainly, specific departmental requirements or offerings could have changed since the 1995 to 1997 study period. However, given our introduction’s review of historically stable faculty and student

26 Teaching of Psychology
evaluations of courses, some of our institutional details may be dated, but the underlying issues are not out of date.

Course Types and Categories

Our analyses hinge on distinctions between psychology course types. Depending on their objectives, earlier investigators used various course typologies, organized under various labels (e.g., Perlman & McCann, 1999a, 1999b). We generally followed the precedent of Rajeev et al. (2005), who established four types: human interest, scientific, applied, and methodology.

Human-interest courses are the most popular type at the national level among all undergraduates and include developmental, social, abnormal, and personality (Adelman, 1995). Scientific courses cover material at a more abstract level of analysis and include physiological, cognition, motivation, and learning. Applied courses are those with a practical orientation such as educational, community, or industrial/organizational. Methodology courses have an explicit focus on statistics, measurement, or laboratory and field research methods.

We categorized the first three types—human interest, scientific, and applied—under the rubric of content courses. Regarding statistics, measurement, and research, even if such courses are not identical, their principles and practices are intertwined in important ways. We collectively categorized these several titles under the rubric of method, which according to our dictionary is a systematic procedure, technique, or mode of inquiry employed by or proper to a particular discipline.

Units of Analysis

Sequence ranks: Enrollment as unit of analysis. A critical datum is the position of a particular title in the sequence of a student’s psychology course enrollments. Regardless of a student’s success in a course, the very sequence of recorded enrollments presumably reflects the interplay between personal preferences and program requirements. For this unit of analysis, we coded all enrolled courses—including duplicates—with any letter grade (A to F, P [pass], S [satisfactory]) or action (W [withdraw], R [defer], I [incomplete]). Alternatively, counts of successful course completions (A to C grades only) might better serve other research goals. Here, the point is moot because our records show that over 90% of first enrollments in method courses resulted in successful completions.

For a given student, the semester order of enrollments established course sequence ranks. The first psychology course to appear on a transcript received the rank of 1.00. In a subsequent semester, the second course got the rank of 2.00, and so on. If two or more courses occurred in the same semester, we assigned the average rank for their positions. Every psychology course in a transcript received a rank. However, for the current analyses we set aside relatively rare early psychology enrollments involving seminars, practica, teaching, writing (grammar and composition), independent readings, and supervised research.

Preferences: Choice as unit of analysis. Another way of demonstrating the interplay between personal preferences and program requirements is to tally the types of content courses first enrolled by undergraduates. As stated earlier, where students presumably have a choice, human-interest offerings seem more appealing than other types (Adelman, 1995). To test this proposition, we noted the type of the first content course enrollment in each of the 784 transcripts. When one or more titles of a given type occurred in a semester, we coded choices as (a) human interest (including applied for this count), (b) scientific, or (c) both.

Results

In terms of the sequence rank analysis, we already identified several reasons to expect that content courses would precede method courses in students’ transcripts. However, of most interest is a possible statistical interaction between programs and the interval between first content and first method enrollments. That is, which program or programs would reveal relatively delayed enrollments in method courses? Regarding the percentage analysis, interaction is also the important issue. Do human-interest psychology courses have precedence at all schools?

Sequence Ranks

We entered the enrollment ranks of the 784 first content courses and 784 first method courses in a mixed design 4 × 2 (× 2) ANOVA, with program as the first factor, student sex as the second factor, and course category (a repeated measure) as the third factor. There was a main effect for course category, with content courses occurring earlier (M = 3.00, SD = 0.96) than method courses (M = 4.69, SD = 2.26), F(1, 776) = 617.55, p < .01. Because sex interacted with neither the program nor the course category variables, we dropped this factor from consideration.

Table 1 shows the average sequence ranks for first enrollments in course categories in relation to programs. The table indicates a marked interaction based on these variables: comparatively small differences in first content course ranks were accompanied by comparatively large differences in first method course ranks, F(3, 776) = 96.35, p < .01. UCIB majors encountered first method courses quite early in their careers (M = 3.59), whereas that experience came quite late for those at UHM (M = 8.12), with IUPUI and IUB in between.

Simple effects tests. Regarding a separate analysis of content courses alone, there were telling differences in the average ranks of first content courses enrolled by students in the four programs. Table 1 indicates that students facing a two-course introductory requirement (IUPUI and IUB) encountered their first content courses later compared to counterparts at UHM and UCB, F(3, 783) = 148.02, p < .01. Indeed, Tukey least significant difference (LSD) tests indicated that the four means in the content column were all different from one another at the .05 level. Regarding a separate analysis of
method courses alone, there were marked program differences in first enrollments, \( F(3, 783) = 128.98, p < .01 \). Tukey LSD tests indicated that the four means in the method column were all different from one another at the .01 level.

*Content and method sequence rank correlations.* For an additional way to display sequence rank information, we calculated the correlation coefficient between students’ first content enrollment and first method enrollment separately for the four programs. The right column in Table 1 indicates effectively no correlation for UHM, but significant negative correlations for IUPUI, IUB, and UCB.

We take these results to mean that, at IUPUI, IUB, and UCB, there were mechanical trade-offs between desire to enroll in a content course and pressure to enroll in a method course. Although students could enroll in any of several courses in a given term, commitment of, say, 3 credit hours to one psychology category (content or method) precluded commitment of the same 3 credit hours to the other psychology category. In line with our hypothesis, the precedence of one category meant a delay for the other. However, at UHM, where there was less pressure to enroll in a method course, trade-offs were not urgent, and students delayed method enrollments in favor of content courses.

**Preferences**

The course mapping technique also revealed differences in the type of content course first enrolled in the various programs. In three departments—UHM, IUPUI, and IUB—students could, to a large extent, choose among early human-interest and scientific content offerings. Table 2 shows that, in these three programs, there was an extremely high likelihood of enrolling in a human-interest course first. (With the human-interest and “both” categories combined, the rate was 95.7%) On the contrary, UCB was the sole program that required a particular scientific (biological) content course, and students there showed a strong tendency to enroll in that course first. (With the scientific and “both” categories combined, the rate was 74.1%.) In Table 2, the contrast across the four programs was marked, \( \chi^2(6, N = 784) = 338.80, p < .01 \). That is, overall there were strong program differences in first content course enrollments.

**Discussion**

Our goals included neither descriptions of a representative set of undergraduate programs nor a chronicle of contemporary curriculum developments. Instead, we capitalized on existing departmental differences at a given point in time, with the aim of connecting various curricular requirements with empirical patterns of course enrollments. We are mindful that many different random variables may have been in effect across the four schools included here. Still, it is tempting to suggest particular antecedents for the striking patterns seen in Tables 1 and 2.

**Sequence Ranks**

majors at UHM enrolled in their first method course later in their careers than did students in the other three departments. A difference among these programs is that UHM required only one method course to satisfy major requirements and all other programs demanded two or more. Confronted with only a single method course, Manoa students may not have perceived pressure for its timely completion. We believe that the UHM students’ actions provide additional evidence for the general tendency among psychology majors to favor content over method course work. A supplementary observation from IUPUI supports this position.

IUPUI departmental recommendations for course sequences were explicit in the 1993–1995 *School of Science Bulletin*. A printed plan of study showed students how to arrange enrollments to cover all department and school requirements within eight semesters. Based on the course mapping procedure described previously, the department recommended an idealized sequence rank of 4.00 for both the first content and the first method course. However, as shown in Table 1, IUPUI majors completed their first content course earlier (\( M = 3.35 \)) and first method course later (\( M = 5.74 \)) than recommended, \( t(103) = 9.07, p < .01 \).

majors at UCB enrolled in their first method course earlier in their careers than did students in the other three departments. The UCB requirements had the following characteristics that could facilitate early enrollment: a single introductory course and no college-level prerequisites for the first method course. Interestingly, these were also characteristics of the program at UHM, where students enrolled so comparatively late. Recall, however, that UCB required two method courses, which in our view accounts for the difference.

**Table 1. Sequence Rank Analysis of First Enrollments in Content and Method Courses**

<table>
<thead>
<tr>
<th>Program</th>
<th>n</th>
<th>M</th>
<th>SD</th>
<th>M</th>
<th>SD</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>UHM</td>
<td>75</td>
<td>2.55</td>
<td>0.96</td>
<td>8.12</td>
<td>3.08</td>
<td>-.05</td>
</tr>
<tr>
<td>IUPUI</td>
<td>104</td>
<td>3.35</td>
<td>0.99</td>
<td>5.74</td>
<td>2.10</td>
<td>-.44*</td>
</tr>
<tr>
<td>IUB</td>
<td>354</td>
<td>3.52</td>
<td>0.67</td>
<td>4.44</td>
<td>1.61</td>
<td>-.36*</td>
</tr>
<tr>
<td>UCB</td>
<td>251</td>
<td>2.25</td>
<td>0.73</td>
<td>3.59</td>
<td>1.55</td>
<td>-.44*</td>
</tr>
</tbody>
</table>

*Note. Values in the M columns are average sequence ranks. The sequence rank of 1 represents the first psychology course enrollment. Higher numbers reflect later positions in students’ patterns of course enrollments. UHM = University of Hawaii Manoa; IUPUI = Indiana University–Purdue University Indianapolis; IUB = Indiana University Bloomington; UCB = University of Colorado Boulder. *p < .01.

**Table 2. Percentage Analysis of First Enrollments in Human Interest and Scientific Psychology Content Courses**

<table>
<thead>
<tr>
<th>Program</th>
<th>n</th>
<th>Human Interest</th>
<th>Both</th>
<th>Scientific</th>
</tr>
</thead>
<tbody>
<tr>
<td>UHM</td>
<td>75</td>
<td>89.3</td>
<td>8.0</td>
<td>2.7</td>
</tr>
<tr>
<td>IUPUI</td>
<td>104</td>
<td>92.3</td>
<td>3.9</td>
<td>3.9</td>
</tr>
<tr>
<td>IUB</td>
<td>354</td>
<td>86.7</td>
<td>8.5</td>
<td>4.8</td>
</tr>
<tr>
<td>UCB</td>
<td>251</td>
<td>25.9</td>
<td>18.7</td>
<td>55.4</td>
</tr>
</tbody>
</table>

*Note. UHM = University of Hawaii Manoa; IUPUI = Indiana University–Purdue University Indianapolis; IUB = Indiana University Bloomington; UCB = University of Colorado Boulder.*
Finally, it is worth noting again the pattern of content and method correlations in Table 1. Negative correlations at IUPUI, IUB, and UCB seem to reflect a competition among enrollment choices. Although we assume a widespread urge to study human-interest material first and foremost, students in those programs somewhat delayed that gratification in deference to the completion of required quantitative courses. In contrast, at UHM, students indulged in early content courses because demands to complete method requirements were lower.

Preferences

Table 2 provides further support for the proposition that—when opportunities allow—psychology majors choose intrinsically interesting courses. At UHM, IUPUI, and IUB, most first content course enrollments were of the human-interest type, including clinical, developmental, and social titles. However, if there is any special merit in giving students early exposure to a scientific psychology content course, UCB shows the path. There, the requirement of a biological psychology course apparently promoted early student involvement. (Nevertheless, many students at UCB managed a first enrollment in a human-interest course.)

Conclusions

This article focused on the dimensions of student course preferences and departmental course requirements. Our course mapping procedure documented the importance of both factors in the early education of psychology undergraduates. Students’ typical interests are clearly revealed in majors’ leanings toward human-interest courses (Table 2), and the influences of program requirements are quite obvious in interdepartmental variability of first method course enrollments (Table 1).

In terms of advising and administration, we see no need for special steps where human-interest courses are concerned; students will find their way into these offerings. On the contrary, educators should take care to ensure early exposure to method courses. Based on the literature and our findings, we make the following recommendations. First, give students reasons to enroll relatively early by (a) setting total method course requirements at two or more and (b) providing instruction and advice on the presumed utility of statistical, research, and technical skills in mastering other types of courses and in the job market. Second, require only a single introductory course. Third, offer (or require) a lower division method course with no college-level prerequisites. Finally, hierarchically structure the major so that some advanced titles require the prior completion of basic method courses.

Although straightforward and even commonsensical, these recommendations are pertinent. As stated earlier, we know that interdepartmental variability in undergraduate research opportunities still exists nationally (Perlman & McCann, 1999b, 2003). Our Tables 1 and 2 show that variations in psychology program requirements can produce striking differences in students’ early educational outcomes. Courses of a certain content type have inherent drawing power, but timely involvement in method courses may call for a bit more regulation.

References


Notes

1. Some of these data were presented as a poster at the Third Annual Hawaii International Conference on Social Sciences in Honolulu in July 2004.

2. For access to transcripts, we extend thanks to Lou McClelland, Director of Institutional Analysis, University of Colorado, and to the staff of the office of psychology advising at Indiana University.

3. Send correspondence to Joan B. Lauer, Department of Psychology, LD124, Indiana University–Purdue University Indianapolis, 402 North Blackford St., Indianapolis, IN 46202–3275; e-mail: jlauer@iupui.edu.
Teaching Statistics to a Student who is Blind
William E. Gibson and Chris Darron

Teaching of Psychology 1999 26: 130
DOI: 10.1207/s15328023top2602_13

The online version of this article can be found at:
http://top.sagepub.com/content/26/2/130

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Society for the Teaching of Psychology

Additional services and information for Teaching of Psychology can be found at:

Email Alerts: http://top.sagepub.com/cgi/alerts

Subscriptions: http://top.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Apr 1, 1999

What is This?
benefits cited by McDade (1995) are additional advantages for case-based teaching.

References


Notes

1. An earlier report of this work was presented at the North Carolina Psychological Association meeting, High Point, NC, May 1996.
2. Send correspondence to Patrick Cabe, Department of Psychology, University of North Carolina at Pembroke, P.O. Box 1510, Pembroke, NC 28372–1510; e-mail: cabep@sasette.uncp.edu.

Teaching Statistics to a Student Who Is Blind

William E. Gibson
Chris Darron
Northern Arizona University

In this article, we describe the use of inexpensive, easily constructed, low-tech teaching devices constructed from cardboard and modeling clay for teaching statistics to a student who is blind. The student haptically explored these manipulatives and achieved a high level of conceptual understanding.

To understand statistics, students must develop the visual and spatial skills necessary for successful visualization and manipulation of numerical data. Teaching statistics to students who are visually impaired is a unique challenge because instructors must find ways to convey the spatial and visual features of data sets and sampling distributions. The somatosensory system is an ideal substitute sensory channel that can be used to convey information in the form of tactile pictures to individuals with visual impairments (Heller, 1991).

Meehan, Hoffert, and Hoffert (1993) recently discussed several useful strategies and resources for teaching statistics to students with visual impairments. For example, a commercially available raised line drawing kit can be used to sketch normal distributions that blind students can feel. Braille versions of distribution tables of z, t, and F are also commercially available. Along with these tactile displays, students can also make use of braille typewriters, tape recorders, computers, and talking calculators to learn statistics.

However, inexpensive and less technically sophisticated devices have been employed to teach both blind and sighted school children mathematical and scientific concepts. Materials such as Styrofoam™, plastic coated wire, macramé twine, and papier-mâché are used to fabricate three-dimensional models and tactile pictures (Bishop, 1996).

We experienced a unique opportunity to develop inexpensive teaching materials for a statistics student who was blind. The student attended class lectures and received daily individual tutoring from a graduate student, worked with a sighted undergraduate colleague who was taking the same class, and used a talking calculator to work the problem sets. Because the problem sets were not in Braille format, the tutor read the numerical data to the student, verbally described the formulas needed to perform the necessary mathematical calculations, and provided the critical values needed for testing against the null hypothesis when inferential statistical tests were used.

Teaching Devices

To enable the student to learn both descriptive and inferential statistical concepts, we created a series of low-tech devices utilizing cardboard and modeling clay and used them during the tutoring sessions. The student touched these manipulatives and haptically explored the contents of the tactile displays (graphs and charts) as a means of visualizing them.

The normal curve, portrayed with modeling clay to create a bell-shaped distribution, demonstrated an array of scores. Vertical strands of clay depicted the locations of the mean and z values and conveyed the principle of areas within the curve from the mean out to a particular z value.

As a means of demonstrating correlations, small pieces of clay, attached to cardboard panels, represented data points within a scatter plot. By touching the array, the student was able to differentiate positive from negative correlations and strong from weak correlations. A strand of clay running through the scatter plot depicted the linear regression line. The student felt the array to perceive the regression line as a floating mean for predicting scores.

Bar graphs made from cardboard illustrated the basic elements of independent groups and matched-group t tests and the one-way ANOVA. Bars of different heights depicted means for a control group and an experimental group. Similarly, cardboard bar graphs reflected differences in mean scores for a group of participants, before and after some manipulation. A set of three bar graphs illustrated data sets that required analysis with an ANOVA. Using the heights of
these bar graphs to portray significant (staircase patterns) or nonsignificant differences, the student developed an understanding of experimental methodology and the use of statistical procedures to evaluate results.

The student successfully learned the material and passed each of the four examinations. The student took oral exams and performed the mathematical calculations to derive necessary statistical values.

Conclusions

Based on this experience with one motivated student who is blind, we believe strongly that students who are visually impaired can learn statistics in much the same way as sighted students. Although many sophisticated teaching devices are currently available that students with visual disabilities can use, it is also possible to fabricate a variety of tactile displays using inexpensive materials that provide a unique type of hands-on learning.

References


Notes

1. Portions of this article appeared as a poster presentation at the Southwestern Psychological Association meetings in Ft. Worth, TX, April 17–19, 1997.

2. We thank Randolph A. Smith and three anonymous reviewers for their many helpful comments concerning this article.

3. Send correspondence to William E. Gibson, Department of Psychology, Northern Arizona University, Flagstaff, AZ 86001; e-mail: william.gibson@nau.edu.

Searching for a Common Core: An Examination of Human Sexuality Textbook References

Matthew Hogben
State University of New York Health Science Center at Brooklyn

Mark G. Hartlaub
Lauren R. Wisely
State University of New York, Albany

We examined a sample of 15 human sexuality textbooks published between 1989 and 1997 to determine the overlap in references among textbooks. We found that fewer than 20% of references were cited in more than 1 text and that less than 1% were cited in 11 or more of the 15 textbooks. We discuss possible reasons for these findings, including implications for the state of human sexuality research.

The majority of research studies in psychology constitute incremental additions to the general body of psychological knowledge. That is, researchers typically compile a body of research that is extended study by study. Given that this principle of scientific advancement applies to the psychological study of human sexuality, textbooks representing the psychology of human sexuality should reflect this body of research to students. That is, the authors of these textbooks should cite from the same sources of information.

To test this hypothesis, we reviewed reference lists in human sexuality textbooks. Shima (1977) conducted a similar review in which he described the articles and authors most frequently cited in 18 supplementary readers designed for introductory psychology classes. Shima found articles chosen for inclusion in the readers to be highly idiosyncratic. Indeed, slightly over 75% of all the articles cited appeared in only 1 of the 18 readers. He also found that only eight articles appeared in 4 or more of the 18 readers (22%) he reviewed.

Gorenflo and McConnell (1991) compiled a list of the most frequently cited journal articles and authors in introductory psychology textbooks. They found that several hundred articles appeared in at least 5 of the 24 textbooks they reviewed and concluded, "therefore, that there is less idiosyncrasy in introductory texts than in readers" (p. 10). They also examined the date of publication of the most commonly cited articles, estimating that it took approximately 20 years for an article to be considered a classic by most introductory psychology textbook authors.

The several hundred, however, does not comprise a large core of common citations as a proportion of the total number of references. As a reviewer of this article observed, the number Gorenflo and McConnell (1991) cited as common to 5 or more of the 24 textbooks they surveyed comprised approximately 3% of the total number of references.

Introductory texts comprise multiple-subject areas, and this heterogeneity of subject material may contribute to lack of overlap. For example, three authors with pairwise reference overlaps of 70% should produce 49% total overlap (.70^2) on a basis of a random distribution of differences across categories. The subject matter of human sexuality texts subject matter is by definition less heterogeneous. Thus, decrements in overlap due to compounded citation differences across categories is not as likely. A core of several hundred references should comprise well over 3% of the total.

Method

We chose a convenience sample of 15 human sexuality textbooks for study. Included within this sample (see Appendix A) were the three best-selling textbooks as of early 1996 (according to a publisher who wished to remain anonymous). Coders generated a reference list consisting of all references found in each of the 15 textbooks. Coders entered each reference, followed by the code (1 to 15) for those textbooks in which the reference appeared.
An Active Learning Approach to Teaching Statistics

Beverly Dolinsky
Endicott College

In this article, I provide practical suggestions on creating a collaborative environment in which to use active learning strategies as the primary method to teach statistics. I describe the use of computer applications and discuss methods of incorporating writing-intensive assignments within a statistics course.

When first teaching statistics, I was a lecturer who emphasized statistical theory. Over time, I found I wanted to shift the focus of the course so students became more involved in their learning. I also wanted students to learn how to interpret the meaning of statistics rather than memorize abstract mathematical concepts. In this article, I provide suggestions to create a collaborative teaching environment in which active learning is the primary method used to teach statistics. The suggestions in this article have come from my own experiences, discussions with colleagues, teaching conferences, and literature on the teaching of psychology (e.g., Becker, 1996; Dunn, 1996; Oswald, 1996; Smith, 1995; Ware & Chastain, 1989, 1991).

Classroom Use of SPSS

For several years I have been moving away from a lecture-based course by refining active learning strategies (as defined in Bonwell & Eison, 1991). One of these active learning strategies is the consistent use of SPSS Base 7.5 for Windows (1997). The course meets in a computer lab and a classroom. I structure the course so that 15 to 30 min of every 75-min session consists of lecture material. The remaining class time is devoted to computer work, group work, verbal and written presentation, and assessment.

I create assignments that require students to discover statistical principles independently. For example, students explore the concepts of central tendency and variation by creating frequency distributions of five ratio variables using data sets provided by SPSS. Students examine their data to estimate the means, medians, modes, and standard deviations. They then calculate measures of central tendency and variation using SPSS. Students describe the statistics’ meanings and decide whether their results support their original estimates.

Teaching the concept of an interaction is much easier and more enjoyable using SPSS. Last year, my students enjoyed making predictions about who watches X-rated movies using data sets provided by SPSS. They were shocked to see that the percentage of people having seen such a movie was much lower than they expected. I immediately took advantage of this teaching opportunity and asked the class why this result might be true. Students began to hypothesize how age and gender might influence this result. After a quick lesson on the crosstab option in SPSS, they were able to see that male, college-age students are much more likely to view X-rated films than older men or women of any age. The remainder of the class period was spent having students develop their own interaction hypotheses. The group exercise allowed students to independently explore complex behaviors that interested them. Their interest in the topic led them to enjoy the assignment and spend more time attempting to master the concept.

Active Learning Strategies

The students spend a significant portion of class time completing inductive reasoning assignments. For example, I introduce the concepts of correlation and regression by showing students three scatterplots that I create prior to class using SPSS data sets. The graphs clearly demonstrate positive, negative, and zero correlations. By examining the graph, students describe the relations between pairs of variables. I ask them how and why the graphs are different. Students write their answers and then present them to the class. This assignment allows an introduction to predictor and criterion variables, scatterplots, types of associations, Pearson correlation, and regression. Following an in depth presentation of these statistical concepts, students then use other data sets to predict additional linear relations.

Other active learning strategies used include course periods devoted to answering open-ended, student-generated questions (e.g., Why is power important? When do you use the ANOVA instead of the t test?). I avoid the tendency to immediately respond to student questions. Instead, I first attempt to have other students answer the question or to provide questions that guide the students to the correct answer. Initially, my students tend to be frustrated when I do not automatically give them a response. They find the questions difficult, and some believe that I am not teaching because I am not giving them the solution. However, over the course of the semester, students begin to enjoy the challenge of collaboratively finding the answers.

There is significant classroom discussion and debate regarding the use, meaning, and value of statistical tests. I frequently give immediate tests to assess mastery of a day’s topic. These tests have little overall bearing on the grade but allow me to adjust the next class period’s lesson.

Writing Intensive Assignments

All homework and class assignments, as well as exams, require written interpretation of data. Some of these writing assignments ask students to interpret the meaning of generated statistical data. For example, I have asked students to describe opinions regarding homosexuality by presenting a variety of frequency and crosstab distributions. Other questions ask students to explore the veracity of a hypothesis by performing the appropriate statistical operations. For example, I have asked them to determine whether women believe more in life after death than men, using data from SPSS. Students always provide written interpretations of the data. Simply calculating the correct answer is not sufficient to achieve a passing grade.
I require a second form of writing assignment called the “one-page press release” (Beins, 1993). Periodically, I ask students to read a preselected refereed journal article. The article includes statistical analyses currently being taught in the class. I ask students to summarize the article into a one-page press release using non-statistical terminology. The assignment helps to develop students’ reading and writing skills as well as their understanding of professional literature.

Another form of writing assignment consists of a learning assessment journal (Qualters & Dolinsky, 1995). The journal assignment asks students to monitor their own learning processes. About every 2 weeks, I assign students a general topic to write about. Topics include their feelings toward statistics, experience with computers, strategies to solve problems, and studying strategies. The journal entries typically range in length from 100 to 250 words. The entries are supposed to be evaluative and demonstrate perceptions of learning in the course. I read and comment on their entries. The purpose of the journal is to allow students to develop self-knowledge of their learning, not only in the statistics course but other courses as well. From my perspective, the journal is also helpful in monitoring student mastery of topics. It allows me to advise students on effective learning strategies and to act as a motivator for students who admit to having frustrations and difficulties.

Conclusions

My goal in creating an active learning environment is to have students become more involved in their learning and develop their critical thinking skills. Is this a better method? Student grades still fit a normal distribution and teaching evaluations remain some of the highest in the college. Student comments, however, are considerably different. Students speak of understanding the material rather than just memorizing facts. They describe feelings of pride and self-confidence at being able to independently solve problems. They also describe insights regarding strengths and weaknesses in their learning styles. I firmly believe these learning achievements were the result of using an active learning approach to teaching statistics.

References


Notes

1. An earlier report of this article was presented at the American Psychological Association meeting, San Francisco, August 1998.
2. Send correspondence to Beverly Dolinsky, Division of Arts and Sciences, Endicott College, 376 Hale Street, Beverly, MA 01915; e-mail: bdolinsk@endicott.edu.

Statistically Lively Uses for Obituaries

Betsy Levonian Morgan
University of Wisconsin–La Crosse

I describe the benefits of using a real data set comprised of local obituaries to illustrate a variety of research issues, such as missing data, outliers, comparing means, and hypothesis testing. The inherent flaws in this type of data set give students hands-on experience with the concerns encountered in analysis of actual data.

Illustrating common principles with hands-on activities (e.g., Weaver, 1992; Zerbolio, 1989) and generalizing underlying principles to new situations (Evans, 1976) are two effective pedagogical approaches to promote the learning of statistical concepts. The activity described in this article combines both approaches by having students analyze a data set they constructed from information contained in obituaries. This activity also addresses several areas of concern to faculty who teach statistics, such as illustrating statistical concepts and developing students’ skills while reducing their fears (Ware & Brewer, 1999). By collecting obituaries over several days, I created a set of real data that students could easily enter into a computer program, analyze, and interpret. I have used this activity with psychology undergraduates designing independent research projects and with secondary education math teachers as part of a continuing education program. All of these students had completed a course in statistics, but most were new to data sets and statistical software. Students’ summary appraisals of the activity were highly positive. Obituaries are rich in detail and represent a wide range of variability. The human-interest factors inherent in obituaries make them salient to students.

The Data Set

Students received a set of obituaries that I photocopied from a local newspaper. They assigned an identification num-
A "Jigsaw Classroom" Technique for Undergraduate Statistics Courses
David V. Perkins and Renee N. Saris
Teaching of Psychology 2001 28: 111
DOI: 10.1207/S15328023TOP2802_09

The online version of this article can be found at:
http://top.sagepub.com/content/28/2/111

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Society for the Teaching of Psychology

Additional services and information for Teaching of Psychology can be found at:

- Email Alerts: http://top.sagepub.com/cgi/alerts
- Subscriptions: http://top.sagepub.com/subscriptions
- Reprints: http://www.sagepub.com/journalsReprints.nav
- Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Apr 1, 2001

What is This?
A “Jigsaw Classroom” Technique for Undergraduate Statistics Courses

David V. Perkins
Renee N. Saris
Ball State University

Undergraduate statistics students vary widely in performance, and many are passive learners. Worksheets (problem sets) help students to be more active and to learn by doing. Working individually, however, students may require too much time to complete worksheets in class, when the instructor is available to help. In the “jigsaw classroom” technique described, a worksheet is divided into 2 to 4 complementary steps that are distributed to different groups of students. Students with the same step complete it together and then collaborate with other classmates to finish the entire worksheet in class. Students reported that this jigsaw technique helped them understand a statistical procedure, used class time efficiently, and increased the variety of learning experiences available in this challenging course.

Wide variability among students in ability, motivation, and assertiveness challenges effective teaching in the undergraduate statistics course (Conners, Mccown, & Roskos-Ewoldsen, 1998). Variability is especially likely with large class sizes at institutions serving students of diverse backgrounds and abilities. It shows up in test performance, attendance, the effectiveness of note taking during lecture, and when students complete worksheets (problem sets) during class.

Spending class time on worksheets is useful for several reasons. Instructors receive immediate feedback on students’ abilities to apply statistical ideas to sample data, and repeated practice with an application helps many students. More important, completing worksheets helps students to be active rather than passive and to learn by doing (Conners et al., 1998). Students see the logic of statistics at work (e.g., how large variability among raw scores affects the computed values of a standard deviation or an F ratio), and their success in mastering specific course material is immediate and tangible, rather than delayed until after a test.

There are two problems with using worksheets in class. We noted that students have uneven ability or readiness to complete worksheets individually. One student may have no trouble, whereas another who is unable to progress on a worksheet may become anxious and frustrated by the apparent ease with which classmates succeed. Less successful students may dread worksheets and avoid trying to work on them at all. A second drawback is that assigning worksheets in class may require too much time because statistical procedures often entail many separate steps (e.g., constructing multiple columns in a table of raw data before inserting summary values into a computational formula).

A solution to both problems comes from adapting Aronson’s (Aronson, Blaney, Stephan, Sikes, & Snapp, 1978) technique of the jigsaw classroom. This technique was created for grade school students to reduce the salience of competition within a classroom, especially a classroom where there exist sharp differences among students in background or ability. A jigsaw procedure lessens the impact of disparities in ability among students by breaking the overall task into pieces and distributing a different piece to each small group of students. Because the group’s success depends on each student, a statistics worksheet structured as a “jigsaw” task becomes a challenge that unites students rather than a contest that divides them. With interdependence among students programmed into the process, fellow students are not competitors but resources for learning (Johnson, Johnson, & Smith, 1991). Jigsaw arrangements also reduce the problem of inadequate class time for worksheets. Instructors can divide many statistical computations (e.g., ANOVA) into two or more subparts that students work on simultaneously before putting them together for the final result. Cooperative work goes faster, making it easier for students to finish a worksheet during class, where the instructor is available to help.

Cooperative learning techniques like the jigsaw method have other classroom benefits (Meyers, 1997). Johnson et al. (1991) reviewed research suggesting that cooperative learning fosters positive attitudes toward the subject of study. In our experience, many students approach the statistics course with dread and anxiety. Incorporating cooperative learning exercises into the course may improve student attitudes.

Cooperative learning methods can also benefit student achievement (Johnson et al., 1991; Slavin, 1983, 1990). Having one or more partners may bring increased individual attention to a less able student, beyond what the instructor is able to provide. Students may see that classmates are also struggling, and together they may feel more comfortable asking for help. Assisting less able classmates may help the more able students learn the material more thoroughly (Cumming, 1983). Additionally, team work experiences can be useful in other challenging learning situations. Working in small teams is common in today’s business and professional settings. Students need to practice interacting with peers on the basis of what each individual already knows, what the group still needs to find out, and how their respective knowledge meshes in producing the final product.
Procedure

Two examples adapted from a commercially published workbook (Pyrczak, 1989) show how the jigsaw process works. The first procedure was a one-way ANOVA with four groups. We put raw data for each group on one of four handouts. Pairs or small groups sitting in adjacent seats all received the same handout and helped one another compute the sample size, sum of the raw scores, sum of the squared raw scores, and sum of squares for one of the four groups. Each handout also included a blank ANOVA table, formulas and instructions for completing this table in collaboration with three classmates (each of whom had one of the other handouts), and a concluding question about whether the null hypothesis of no differences among groups should be rejected. Completing this group exercise with classmates having complementary handouts enabled each student to answer a series of questions on the original ANOVA worksheet relatively quickly.

A second worksheet presented two-way, chi-square tests of independence for three different studies, involving 3 × 3, 2 × 2, and 2 × 3 tables, respectively. We demonstrated the computation and interpretation of chi-square for the first study (involving the expected frequencies for each cell, value of chi-square, degrees of freedom and, using a table from the text, the critical values needed to reject the null hypothesis at various alpha levels). Following discussion of the first example, students received one of two handouts directing them through every step of the chi-square procedure for one of the remaining designs in the worksheet (and giving a partial solution for each step). Each handout also provided the next-to-last step (the full computational equation for chi-square) for the remaining design in the worksheet. Thus, one group of students received step-by-step instructions and partial solutions for the second design, plus a nearly complete solution for the third design. The remaining students received the step-by-step information for the third design, and the almost-complete solution for the second design. The final instruction on both handouts directed students to seek out a classmate with the complementary handout. Pairs of students finished the entire three-problem worksheet in only a few minutes.

Evaluation

Students rated the benefits of these exercises using a 5-point Likert scale, with anchors at 1 (not at all useful) and 5 (very useful). Specific items (with corresponding benefits in parentheses) were: (a) getting help, (b) giving help, and (c) working with classmates (increasing contact, cooperation, and support; helping students become collaborative learners); (d) providing an alternative to lecture (increasing the variety of learning experiences); (e) saving time (using the class time devoted to worksheets more efficiently); and (f) understanding the statistical procedure (increasing student achievement). Percentages of students rating the usefulness of jigsaw techniques as either 4 or 5 (the most positive choices) were: working with others, 55%; getting help from others, 55%; giving help to others, 67%; understanding the statistical procedure, 66%; saving time in the completion of worksheets, 84%; and as an alternative to lecture, 88%.

Two kinds of evidence are consistent with positive student attitudes as a benefit of using jigsaw techniques. First, in response to the question “Would you recommend using this group jigsaw exercise again?” 21 of 24 (88%) students answered yes for the first exercise, and 18 of 21 (85.7%) answered yes for the second exercise. Second, mean ratings of the instructor on standardized student course evaluations improved significantly between the year before the jigsaw technique was developed and the year it was first used, $t(53) = 3.56, p = .001$.

We also assessed student achievement by comparing exam performance for students who took the course in years prior to using the jigsaw technique and in years when the technique was used. Although exams were mostly conceptual and few items directly measured the concrete learning addressed by these jigsaw procedures, we identified four items used both before and after the jigsaw technique was implemented that were specifically relevant (requiring correct calculations of mean square, $F$ ratio, and chi-square expected-frequency values). Student performance was very good on these items, ranging between 74% and 97% correct across items and years. Performance by prejigsaw and postjigsaw students did not differ on these specific items. However, overall scores on the exam covering ANOVA and chi-square improved from a mean of 63% correct in prejigsaw classes to 68% correct for postjigsaw classes, $t(161) = –2.4, p = .017$. Thus, these jigsaw exercises maintained students’ relatively high performance on computational exam items, and their general benefits (increased understanding, improved morale, class time saved) were associated with improved learning overall.

Discussion

Student perceptions of the jigsaw procedure were very positive, especially as an alternative learning experience. Instructors who are reluctant to displace time spent on inactive techniques (like lecturing) may note that a jigsaw exercise saves time over doing worksheets individually and that in our classes exam performance was as good or better for jigsaw versus nonjigsaw students. There may be little reason to maximize the time spent on inactive techniques like lecturing if students who collaborate actively during class sessions value these experiences and learn just as much. The improvement in student ratings of the instructor is important in suggesting that students appreciated the instructor’s efforts to accommodate the course to different learning styles and had more positive attitudes about their experience in learning statistics.

Regarding other purported benefits, it is interesting that students saw the exercises as somewhat more useful for practical purposes (understanding the statistical procedure, saving time) than for the interpersonal benefits of working with others, getting help, and giving help. Although nearly all students appreciated the technique as a time-saver and a change of pace from lectures, the social demands of a jigsaw experience may not appeal to every student. Some may prefer to complete all
parts of a worksheet at their own pace; others may feel pressured when classmates are waiting on them to finish.

We have several other suggestions for those using jigsaw techniques. Because individual students spend most of their time on only one piece of the worksheet, the instructor should discuss the jigsaw exercise following completion to clarify students’ understanding of the overall statistical design. Second, jigsaw techniques are not designed for learning a new skill on the first attempt (Aronson et al., 1978). Instructors should use this technique after determining that students understand how to complete their pieces of the overall worksheet. Students learn these steps quickly if the instructor works through a complete example that is similar to the procedures needed for the jigsaw problem.

A jigsaw technique based on paper-and-pencil worksheets complements the use of computerized statistics packages by helping students understand how to derive a test statistic by hand using raw data. To obtain the collaborative benefits a jigsaw experience provides, instructors using software packages could divide computer-based assignments into complementary pieces that students assemble and analyze (e.g., students create, match, and execute the different data and syntax files required by contrasting statistical procedures).

In sum, different students have different learning styles and too many are passive learners, especially in classes like statistics. By making each student part of the solution, the jigsaw technique blurs the distinction between students who know and students who do not yet know, requires all students to make active responses, and moves away from the experience of learning as a solitary activity that is detached from the social context. The technique also helps students appreciate that one of the best ways to learn is to teach others (Webb, 1992).

References


Note

Send correspondence to David V. Perkins, Department of Psychological Science, Ball State University, Muncie, IN 47306; e-mail: dperkins@gw.bsu.edu.
Personal Ad Content Analysis Teaches Statistical Applications
D. W. Rajeeck
Teaching of Psychology 2002 29: 119
DOI: 10.1207/S15328023TOP2902_08

The online version of this article can be found at:
http://top.sagepub.com/content/29/2/119

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Society for the Teaching of Psychology

Additional services and information for Teaching of Psychology can be found at:
Email Alerts: http://top.sagepub.com/cgi/alerts
Subscriptions: http://top.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Apr 1, 2002
What is This?
Personal Ad Content Analysis Teaches Statistical Applications

D. W. Rajecki
Indiana University–Purdue University Indianapolis

Students analyzed newspaper personal advertisements for the age preferences of younger and older male and female writers. Several applications of the F statistic revealed between-group preferences for partners of relatively different ages. Many applications of the r statistic revealed within-group preferences for partners of absolutely similar ages. Students communicated this formal statistical information in a written report. Subsequently, most students rated the personal ads as interesting and as giving the impression of working with real people. Success on final examination computational items (including naming degrees of freedom and probabilities) exceeded 90%. A separate quantitative assessment revealed a reliable increase in mastery of r and F statistical notation and concepts from before to after the project.

Personal advertisements yield convenient samples of inherently interesting societal raw data. Researchers gather such material from national tabloids (Harrison & Saeed, 1977), local newspapers (Rajecki, Bledsoe, & Rasmussen, 1991), and the Internet (Matthews, 1999). Ad content analysis provides beginning students with useful exposure to basic statistical applications and procedures. Identifying ads as an engaging pedagogical tool contributes to the growing literature on methods and issues in statistical instruction (cf. Friedrich, Buday, & Kerr, 2000; Ware & Brewer, 1999; Ware & Johnson, 2000).

My technique is based on a recent study by Rasmussen et al. (1998) that examined the partner age preferences of ad writers whose own ages ranged from the 20s to 50s. Those researchers reported as many as 30 separate correlation coefficients (r) indicating that across the life span male and female advertisers consistently sought others of roughly similar age. At the same time, F tests indicated that women generally preferred somewhat older partners, whereas men preferred somewhat younger partners. Finally, in later segments of the life span, writers of both sexes tended to prefer progressively younger partners.

For my purposes, I scaled down the Rasmussen et al. (1998) design to give undergraduates meaningful and manageable opportunities to work extensively with the r and F statistics. I created opportunities for mechanical practice with computer data entry and computation. Writing assignments (including tables) provided experience with the proper communication of statistical information such as degrees of freedom and probability estimates.

In the classroom, this project can certainly be couched in theoretical terms. For example, are people's mate preferences better explained by concepts from evolutionary theory (Kenrick & Keefe, 1992) or social exchange theory (Rasmussen et al., 1998)? Alternatively, students can empirically address a number of interesting sets of questions. For example:

1. Compared with one another, do men and women in different age decades (say, the 20s, 30s, or 40s) have different partner age preferences?
2. All else being equal, do men and women prefer partners with ages similar to their own ages?

The first set of questions can be answered by evaluating group mean differences with the F statistic. The second set of questions can be answered by showing the strength of association between own and partner ages with the r statistic. In my classes, I gave students practice with both types of tests: They calculated two separate ANOVAs and 16 separate correlation coefficients.
Method

Participants

With a few exceptions, undergraduates in the local Introductory Laboratory in Psychology course complete a statistics prerequisite, and most are psychology majors. Laboratory sections admit a maximum of 23 students each, but they usually enroll fewer. I taught six of these sections during three semesters in 1999 and 2000. Sample sizes involved in specific assessments are provided subsequently.

Ad Samples

I assembled sets of heterosexual personal ads clipped from the pages of a metropolitan newspaper published in the mid-1990s. The initial requirement for inclusion was that the ad needed to state three explicit ages in years: that of the writer (own) and both an upper and lower limit for the potential reader (other). Note that for the sake of the exercise, I picked ads that would reveal the general trends and effects found in the Rasmussen et al. (1998) article. There were 25 ads in each of six separate sets based on men versus women writers in three age decades: 20s, 30s, or 40s.

Project Requirements

The project final assignment was a paper written in American Psychological Association style that included a method section, a results section, and tables for F and r results. The due date for the paper was 3 weeks after the distribution of the ads. Two weeks after the original submission of the papers, I provided explicit, written criticisms. These critiques listed points deducted for omitted information, calculation errors, and improper statistical statements. To recoup points, students submitted revised papers, due 2 weeks later.

Data Management and Statistical Analyses

Derived scores. I distributed photocopied sets of the 150 ads and a supplemental handout containing recording forms, instructions, and examples of tables. A first task captured the age information from each ad. On six separate forms reflecting writer gender and age decade, students filled in columns for the writer’s age (own) and the lower and upper ages stipulated for the reader (other).

Then, in a stepwise process, students derived an index that allowed for an efficient, composite expression of relative age preferences. For a given ad, they added the lower and upper year limits stipulated for the other and divided the sum by two. This strategy established the midpoint of the other’s age range. Students entered these whole or fractional midpoints in another column on the form. Next, for each ad they subtracted the writer’s age from the corresponding midpoint age. This final result was labeled the difference between the midpoint and own age (DMOW). The DMOW procedure reduced three age values (own, other lower, and other upper) to a single derived score. Resulting negative DMOWs revealed net preferences for younger partners, and positive DMOWs indicated net preferences for older partners.

ANOVA (F). The preceding procedure led to six average DMOWs based on gender and age decade. These means were useful to answer the first set of questions posed in the Introduction. The assigned paper required two separate 1 × 3 ANOVAs—one for the three men’s DMOWs and one for the three women’s DMOWs. Students presented mean DMOWs and their attendant F statements in a required table.

Correlation (r). The paper also required an array of correlations (r) for a comparison of writers’ absolute ages with the absolute age limits stipulated for the other. These coefficients were useful to answer the second set of questions posed in the Introduction. Students calculated a total of 16 separate coefficients. Two of these rs were based on the overall sample of 75 ads written by men: One compared own ages with corresponding lower end stipulations, and the other compared own ages with corresponding upper end stipulations. Two more of the 16 rs dealt with parallel comparisons for the overall sample of 75 ads written by women. The remaining 12 rs involved the same type of age comparisons, but were based on the disaggregation of both gender samples into three subsets of 25 ads each for 20-, 30-, and 40-year-olds. Students presented the 16 coefficients and their attendant p statements in a second required table.

In class, individuals carried out F and r calculations using a computer program called Student Statistician. This downloaded package accompanied Psychological Research: Methods for Discovery and Validation (Vadum & Rankin, 1998).

Assessment

Subjective Measures

Late in the first two of the three semesters in the study period, 48 students volunteered to fill out a 10-item anonymous attitude questionnaire concerning features of the personal ads, the project, and the statistical package. The essential content of these items appears in Table 1. A 7-point bipolar scale with endpoints relevant to particular content accompanied each item. For example, one of these scales ranged from 1 (not at all) to 7 (definitely). For all questions, the higher the obtained number, the more favorable the attitude.

Table 1 presents average responses to the attitude questionnaire’s 10 items. All the means in the table are above the scale mathematical midpoint of 4, indicating generally positive evaluations by students. Another way to establish the clearly positive nature of students’ judgments is to note that the 95% confidence intervals for ratings on all items fall above the scale midpoint. Finally, the percentage column in Table 1 indicates the number of students who marked the 5, 6, or 7 scoring levels on the 7-point scales. The majority of these individuals expressed endorsement of the personal ads, the project, and the statistical package. Note that students who responded positively to the r item were also likely to endorse the F item, r(46) = .53, p < .01.
During final examinations in these two semesters, 52 students completed $r$ and $F$ computations based on unfamiliar data sets. These problems included naming degrees of freedom and probability levels. The correct answer rate for the $F$ exam item was 95.1%, and the corresponding rate for the $r$ item was 90.4%.

However, final examination performance alone cannot document acquisition of knowledge. To that end, I later devised a six-item questionnaire to tap understanding of the utilization, interpretation, and expression of $F$ and $r$ statistics.

At the beginning of the third semester in the study period, I administered the statistical questionnaire to students in my two laboratory sections. For control purposes, I also administered the instrument at the first meeting of my concurrent social psychology lecture course, which had many nonmajors and no statistics prerequisite. (For the assessment, there was no overlap of laboratory and lecture students.) Finally, in that same early week, eight PhD faculty members consented to a one-time completion of the statistical questionnaire. Later, near the end of the term (and with no announcement), the lecture and the laboratory students completed the statistical questionnaire a second time. Analyses of the undergraduate data involved only respondents present at both testing sessions (lecture $n = 28$, lab $n = 23$).

The material in the questionnaire proved to be challenging. The maximum score was 6, but four of the eight PhD respondents got at least one answer wrong, yielding a faculty average score of 5.38 ($SD = 0.74$). As expected, first test scores for the undergraduate samples were much lower. The social psychology lecture students produced an average of 0.39 ($SD = 0.74$), whereas their more sophisticated laboratory counterparts scored a mean of 1.22 ($SD = 1.09$). However, by the second testing session, students’ scores improved. The posttest mean in the control social psychology lecture course rose to 1.21 ($SD = 1.32$), and an even more striking shift emerged in the lab students’ average of 3.17 ($SD = 1.47$).

A mixed-factor ANOVA revealed a main effect for the consistently superior performance of the laboratory students, $F(1, 49) = 29.95$, $p < .01$. Furthermore, the total change between the two test sessions produced another reliable main effect, $F(1, 49) = 43.78$, $p < .01$. Finally, improvement between the two test sessions was higher among lab students compared with lecture students as revealed by the interaction term, $F(1, 49) = 7.31$, $p < .01$. That is, the lecture students showed an average shift of 0.82 units compared with 1.95 for those in the laboratory course. Put another way, from the first to second testing 19 of 23 laboratory students showed some positive score shift, compared with 13 of 28 lecture students, $\chi^2(1, N = 51) = 7.07$, $p < .02$.

### Objective Measures

<table>
<thead>
<tr>
<th>Item</th>
<th>$M$</th>
<th>$SD$</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personal ad features</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Personal ads are interesting subject matter?</td>
<td>5.67</td>
<td>1.31</td>
<td>88</td>
</tr>
<tr>
<td>Impression of working with real people?</td>
<td>5.50</td>
<td>1.43</td>
<td>81</td>
</tr>
<tr>
<td>Project features</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First opportunity to work with data management?</td>
<td>6.06</td>
<td>1.56</td>
<td>83</td>
</tr>
<tr>
<td>Project increased your understanding of $r$?</td>
<td>5.23</td>
<td>1.21</td>
<td>77</td>
</tr>
<tr>
<td>Project increased your understanding of $F$?</td>
<td>4.67</td>
<td>1.57</td>
<td>63</td>
</tr>
<tr>
<td>Multiple statistics and tables challenging?</td>
<td>5.00</td>
<td>1.29</td>
<td>73</td>
</tr>
<tr>
<td>Opportunity for revision increased learning?</td>
<td>5.44</td>
<td>1.61</td>
<td>81</td>
</tr>
<tr>
<td>Learning in lab different from lecture course?</td>
<td>4.75</td>
<td>1.39</td>
<td>63</td>
</tr>
<tr>
<td>Statistical package features</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Statistical package easy to use?</td>
<td>6.67</td>
<td>0.60</td>
<td>100</td>
</tr>
<tr>
<td>Statistical output easy to understand?</td>
<td>5.21</td>
<td>1.62</td>
<td>75</td>
</tr>
</tbody>
</table>

**Note.** Particular scale anchors were relevant to item content. For example, for the first two items the endpoints were not interesting versus interesting and not at all versus definitely. The % column indicates the percentage of students who marked the 5, 6, or 7 scoring levels on the 7-point scale.

**Table 1. Student Ratings of Features of the Personal Ads, the Project, and the Statistical Package**

**Discussion**

The subjective endorsements shown in Table 1 are complemented by the objective assessments. Student success on final examination computational items was over 90%. Regarding evidence for the acquisition of knowledge, results from the statistical questionnaire indicated a reliable positive shift in mastery of technical concepts from the first to second testing. Although I am mindful of the limitations of the instrument and the samples, it is nevertheless gratifying that by the end of the exercise the laboratory students were scoring at a level ($M = 3.17$) about 59% of that of the faculty ($M = 5.38$).

Still, certain qualifications are in order. On the one hand, given the field conditions under which I tested, it may not be safe to conclude that the laboratory students scored higher compared with those in the lecture course due only to their experience with the personal ad project. Individuals could have obtained statistical information from books, peers, advisors, or instructors outside of the course. On the other hand, if such extracurricular exposure did take place, the students in this article at least got to use their learning in the context of work with personal ads.

In the semesters reported here, I chose to spend more of students’ time and effort on correlation compared with ANOVA. Beyond my agenda, an instructor who uses per-
sonal ads as a database can develop ways to emphasize his or her own choice of any conventional statistic.

Furthermore, concerning specific content, I restricted attention to age information in ads. However, a glance at any personals section in the newspaper will reveal many other interesting types of ad content. Advertisers frequently touch on matters of appearance, financial resources, personality, good (and bad) habits, and religiosity, among other things. The enterprising instructor will discover many opportunities for various quantitative and qualitative research applications to help students learn statistical concepts.

References


Note

Send correspondence and requests for copies of the personal ad sets, classroom forms, and the attitudinal and statistical questionnaires to D. W. Rajecki, Department of Psychology, LD124, IUPUI, 402 North Blackford Street, Indianapolis, IN 46202–3275; e-mail: drajecki@iupui.edu.
Stat Jingles: To Sing or Not To Sing

Carmen R. Wilson VanVoorhis
University of Wisconsin–La Crosse

I investigated the effect of using music as a mnemonic device to remember statistical facts. Participants were 71 students in 2 sections of Psychological Measurement, a required course for psychology majors. Students in 1 section learned and sang 3 statistical jingles, whereas students in another section read definitions of the statistical concepts aloud. Students who learned the jingles scored significantly higher on related test items. Additionally, ratings of jingle knowledge were significantly correlated with test scores.

Statistics teachers frequently are all too familiar with glazed-over faces and droopy eyelids in response to such topics as standard deviation, the normal distribution, and standard scores. I rarely want students to memorize statistical facts or equations. Rather, I would like them to understand the underlying concepts such that they can apply those concepts in a wide variety of situations. Nevertheless, a few facts exist that I believe are worth memorization, such as the percent of scores within 1, 2, and 3 SDs from the mean in a normal distribution. After teaching Psychological Measurement several semesters with the same disappointing results to my “memorize this” request, I decided to try something new. I wrote three statistical jingles to improve students’ memory for certain statistical facts.

The psychology of memory has long focused on the benefits of mnemonics. Mnemonics generally require a person to concentrate on a piece of information for some time. This increased focus and elaboration can improve the individual’s memory for that information. Even rote mnemonics tend to improve long-term memory for unrelated facts (Carney, Levin, & Levin, 1994). Although most introductory psychology texts extol the use of the method of loci or pegwords, few, if any, discuss the role of music in memory.

Advertisers, however, often use music in the form of jingles to impress their slogans. Wells, Burnett, and Moriarty (as cited in Yalch, 1991) stated that “finger-snapping, toe-tapping songs have tremendous power because they are so memorable” (p. 273). Recent research has demonstrated that after a single exposure, individuals made more correct brand–slogan associations with advertisements using a jingle versus those that did not use a jingle (Yalch, 1991). Slogans are not the only type of message that people can remember when combined with music. McElhinney and Annett (1996) presented two groups of participants with song lyrics. The lyrics were read to one group and sung to another group. The group that heard the lyrics sung remembered more words and showed evidence of greater chunking of the material.

In this study, I investigated the effect of learning three statistical jingles on memory for statistical facts. I hypothesized that students who learned the jingles would earn higher scores on test items related to the material in the jingles than students who did not learn the jingles.

Method

Participants

Students in the experimental section learned three statistical jingles on 2 separate days (see Appendix). The jingles review standard deviation, the normal curve, and standard score distributions. I sang each jingle once alone, and then we sang the jingle twice in unison while looking at the words. Finally, we sang the jingle once in unison without looking at the words. Students in the control section recited definitions aloud, in unison, two times looking at the words and one time without looking at the words.

The dependent variable was an aggregate score of four short-answer test items related to the material in the jingles (coefficient $\alpha = .82$). The items asked students to describe a standard deviation, describe the normal curve, interpret a standard score, and interpret a regression equation. In addition, students in each section rated their knowledge of the each of the three statistical jingles on a scale ranging from 1 (never heard of it) to 7 (could sing it in my sleep).

Results

To establish reliability of the scoring system, I provided grading criteria to a colleague. We both scored each test item, blind to each other’s scores. A Pearson product–moment correlation indicated an interrater reliability based on these scores of .93 ($p < .01$). Students could receive partial points on each item. I computed total points (9 points possible) across the four items for each student and analyzed group differences with a t test for independent samples. Students who learned the jingles earned significantly higher total scores ($M = 5.95$, $SD = 1.91$) than the students who read the definitions aloud ($M = 5.06$, $SD = 1.84$), $t(69) = 2.01$, $p < .05$. In addition, the statistical jingle students’ ratings of their knowledge of the jingles correlated significantly with their total scores on the four items, $r(31) = .37$, $p = .04$.

Discussion

Research findings indicate that jingles can be effective in improving memory compared to nonmusically presented information when there have been few repetitions of the material (Yalch, 1991). In a college classroom, time is frequently of the essence. A challenging goal is to present material in a time
efficient and memorable manner. The statistical jingles accomplished that goal. We spent no more than 5 min on any one of the three jingles, yet students who learned the jingles earned more points on related test items.

In addition, research suggests music increases the chunking of material (McElhinney & Annett, 1996). Two of the jingles have memorable lines early in the jingle. In “Plot, Plot, Plot Your Curve,” the line “The mean is equal to the median and the mode” is easily remembered. Fifty-five percent of students who learned the jingle earned perfect scores on the normal curve item as opposed to 38% of the students in the other section. In “I’m a Standard Deviation,” the line “I estimate the average distance from the mean, across a group of scores” is similarly remembered and recalled with ease. Again, 65% of students who learned the jingle earned a perfect score on the related item as compared to 38% of the students in the other section. An extra credit question on the final exam asked students “What is one thing you learned in this class that you will not forget?” Twenty percent of students mentioned the jingles. The majority of them (67%) quoted one of the previously mentioned lines. Students did not mention information presented later in the jingles, anecdotally suggesting that they had less memory for that information. Future research might address memory for information depending on whether it occurs early or late in the jingle. It is possible that having students sing the jingles on more than one day might have improved overall memory for the song, and future research also might address this issue.

Equally if not more important than the mnemonic properties of the jingles was the effect of singing on class atmosphere. Students actually reported enjoying the jingles. The jingles were a pleasant diversion to ordinary, sometimes even monotonous, statistical lectures. Almost all students willingly sang the jingles with smiles, and they applauded themselves when the song ended. In other words, the jingles helped make the class fun. Encouraging students to view statistics as fun is a worthy goal in and of itself.

References

Appendix
“I’m a Standard Deviation” (to the tune of “I’m a Yankee Doodle Dandy”)
I’m a standard deviation

A standard deviation am I
I estimate the average distance from the mean
Across a group of scores.
I’m a standard deviation
A standard deviation am I
Subtract the mean from each score
Square and add them up
Divide by n, take the root
And that’s a standard deviation

“Plot, Plot, Plot Your Curve” (to the tune of “Row, Row, Row Your Boat”)
Plot, plot, plot your curve
Plot your normal curve
The mean is equal to the median and the mode
Plot your normal curve
Sixty-eight percent of scores
Fall within one
Standard deviation
Above and below the mean
Ninety-six percent of scores
Fall within two
Standard deviations
Above and below the mean
Ninety-nine percent of scores
Fall within three
Standard deviations
Above and below the mean

“The Standard Score Distribution” (to the tune of “She’ll Be Coming Around the Mountain”)
The standard score distribution has a given mean and SD
The standard score distribution has a given mean and SD
The standard score distribution
The standard score distribution has a given mean and SD
The z-score distribution has a mean equal to 0
The z-score distribution has an SD equal to 1
The z-score distribution
The z-score distribution
Has a mean equal to 0 and an SD equal to 1
The T-score distribution has a mean equal to 50
The T-score distribution has an SD equal to 10. …
The sta-9 distribution has a mean equal to 5
The sta-9 distribution has an SD equal to 2. …

Notes
1. A poster based on these data was presented at the annual meeting of the Midwestern Psychological Association, Chicago, May 2000.
2. I thank Betty Miller for scoring the test items and Betsy Morgan and Bart VanVoorhis for their thoughtful comments on earlier versions of this article.
3. Send correspondence to Carmen R. Wilson VanVoorhis, 335 Graff Main Hall, Psychology Department, University of Wisconsin, La Crosse, WI 54601; e-mail: wilson.carm@uwla.edu.
Creating Problems to Solve Problems: An Interactive Teaching Technique for Statistics Courses

David W. Kolar
Christine A. McBride
Mary Washington College

We describe an interactive teaching technique for use in statistics courses. In collaboration, students create and solve practice problems, exchange problems with other students, and form a class resource for studying. The exercise gives students practice computing statistics and writing appropriate conclusions. Most important, students gain an understanding of when to use specific statistical analyses. Feedback from students indicates that they believe the technique is effective.

For many psychology students, statistics courses can provoke more fear than other courses in the major. In fact, many of our students initially question why a psychology major needs to know anything about statistics. Therefore, statistics courses create not only a great deal of anxiety for students, but also a great challenge for instructors.

One way to change the attitudes of students about a course is to move beyond simply lecturing by making the course more interactive (Dolinsky, 2001; Stedman, 1993). This move is particularly needed in statistics, in which constant lecturing about mathematical formulas and concepts is likely to decrease student interest (Dolinsky, 2001). However, we have found that unlike instructor’s manuals for textbooks in courses such as introductory psychology and social psychology, manuals for statistics textbooks are often not as helpful in creating an active learning environment in the classroom. This article describes an interactive exercise for statistics courses that actively engages students in the computational and conceptual aspects of statistics.

Student Created Problems

Beyond learning how to compute specific statistics, students should learn to select appropriate tests to answer research questions (see Ware & Chastain, 1991), interpret test results, and write suitable summaries of those results (see Beins, 1993; Dunn, 1996, 2000). To facilitate the learning of these computational and conceptual skills, we have students (in groups of two or three) create practice problems and data for many of the statistical tests we cover. For example, after covering independent samples t tests in class, one group of students created a word problem comparing the amount of time (in minutes) it takes cats and dogs to consume their daily food. Before creating their problems, we tell students to consider issues such as the type of data needed (e.g., nominal or ratio) and the number of levels of the independent variable. Because students will be solving these problems by hand as well as on the computer, we encourage students to keep their data sets small.

After creating the problems and data, students solve the problems on the back of the paper. Their solutions include the statistical computations as well as a basic conceptual conclusion (e.g., there is not a difference between cats and dogs in the number of minutes it takes them to eat their food). Student groups exchange problems and check each other’s work. In addition to checking for computational errors, students also ensure that the statistical test is appropriate for the research question and data presented. At this time, we consult with the groups and help correct any mistakes. To end the class, we collect all of the word problems and solutions and place them into a box located in a common area that is accessible to students.

The final step in this exercise is for students to use the problems throughout the semester for practice. As the course moves along, students fill the box with problems representing an array of statistical tests. Students randomly select problems during class approximately every other week during the semester and solve them by hand or using a computer statistical package. They read the word problem on the front of the page, attempt to solve it, and then check their work based on the solution on the back of the page. In addition to using this procedure in class, we strongly encourage students to randomly select problems from the box and solve them on their own.

Pedagogical Advantages

There are several pedagogical advantages to having students create and solve word problems in statistics courses. Because students must develop their own problems (they are not allowed to use problems from class or the textbook as a template), they must think about what research questions and data are appropriate for a particular statistical test. When
students first do this exercise, they have difficulty creating appropriate word problems. After doing the exercise a few times, students realize that understanding statistics is more than just plugging numbers into a formula. They focus more on the research questions involved and how statistical formulas are helpful in answering them. By the end of the semester, most students are good at creating appropriate word problems and data.

In addition to the benefits of creating the problems, students gain an understanding of the conditions under which each statistical test is used by randomly choosing problems representing different statistical tests and solving them. One of the most difficult tasks for many statistics students is determining the appropriate statistical test to use given the research question and the data available. When students randomly select word problems from the box during class (and on their own outside of class), they must read the research question, determine the type of data, and make a determination as to the appropriate test to use. By the end of the semester most students are better at selecting tests and the box of problems helps them study for a cumulative final in which they have to choose the appropriate statistic to answer a research question.

Another skill that students learn from this exercise is how to write appropriate conclusions about statistical results. As Beins (1993) noted, students often have trouble translating statistical results into nonstatistical terms. By writing a conclusion about their study rather than simply circling a final answer, students gain a better understanding of what one can and cannot conclude from statistical analyses. The box full of conclusions gives the students many perspectives on how to write a clear, concise summary of what statistical results indicate.

Student Perceptions

We recently surveyed 44 students in a statistics class regarding this interactive exercise. Ninety-three percent of the students either agreed or strongly agreed with the statement “Creating and solving practice problems in class helped me understand some of the computational components of the different tests we learned this semester,” and 64% of the students either agreed or strongly agreed with the statement “Creating and solving practice problems in class helped me understand some of the conceptual components of the different tests we learned this semester.” Only 7% of the students either agreed or strongly agreed with the statement “I thought it was a waste of time to create and solve practice problems in class.” The results of this survey indicate that students generally found this activity helpful in learning both the computational and conceptual aspects of statistics.

Potential Problems and Suggestions for Solving Them

We have encountered a few difficulties in the process of developing this exercise. For example, initially some students did not put much effort into the exercise. They simply replicated a problem from their text or from class and put little thought into creating the data. Although we do not give class credit for this activity, our solution to this problem is to have students write their names on the problems they create. Students take the exercise seriously because they know we, as well as many of their classmates, will be looking at what they create. In addition, we now make it clear that they are not allowed to simply copy a problem from the textbook. A second problem we encountered stemmed from the number of problems students created. Many students became tired of creating problems by the end of the semester. We now limit the number of statistical tests we cover with this exercise to eight or nine per semester. Interestingly, one area in which we have not had very many problems is computational errors. By requiring a second group of students to check for computational errors, we have eliminated the majority of computational mistakes.

One suggestion for improving this exercise is to have students write their conclusions in a scientific format. Currently, we have students focus on writing a short, conceptually coherent conclusion. We are considering placing more emphasis on writing both the statistics and the conceptual conclusion in American Psychological Association style.

Conclusions

Collaborative learning can improve both achievement in and attitudes toward statistics courses (Potthast, 1999). The use of small groups to create and solve statistical problems is a technique that should not only help students understand the conceptual and computational aspects of statistics, but also generate a useful resource of practice problems for students to peruse throughout the semester.

References


Note

Send correspondence to David W. Kolar, Department of Psychology, Mary Washington College, Fredericksburg, VA 22401; e-mail: dkolar@mwc.edu.
tions, perhaps an indication of the merit of including a wide variety of music for students who often come from diverse backgrounds and bring a broad range of experiences with music to the classroom. When asked to identify any pieces they thought were inappropriate or not that relevant, students provided only nine responses. Finally, we asked students in the second semester to suggest any changes to the procedure. The vast majority either did not respond or indicated that no changes should be made.

Conclusions

We identified a substantial number of pieces representing a wide array of musical styles that students considered to be relevant to major subject areas covered in developmental psychology classes. Instructors who include other topics in their classes may also find music that is relevant. For example, musical selections exist that are concerned with applied social policy topics such as foster care, adoption, teenage pregnancy, the effects of segregation, and children’s rights. We believe that faculty could initiate this kind of project in many other kinds of psychology courses as well, although we have not explored this possibility.

Music is an integral part of many students’ lives. However, the extent to which listeners seriously consider some of the meaning or circumstances associated with the lyrics, especially when they depart from the more conventional themes involving romantic relationships (and their dissolution), is unknown. We initiated this activity to encourage students to appreciate how music can inform them about themes and principles of child development. The music provided a favorable way of moving into the day’s classroom activity. We elected to make the music available as a supplement to course requirements. Instructors also may consider implementing the activity, at least on an occasional basis, within the class period. However, because doing so reduces time available for other pedagogical emphases, instructors need to weigh whether playing music within the period is justified or whether it is more useful as a supplement to traditional classroom activities.

References


Note

Send correspondence and requests for complete listings of the musical selections to Marvin W. Daehler, Department of Psychology, University of Massachusetts, Box 37710, Amherst, MA 01003–7710; e-mail: marvin.w.daehler@psych.umass.edu.

Applying the Just-in-Time Teaching Approach to Teaching Statistics

James O. Benedict and Jessica B. Anderton
James Madison University

Just-in-Time Teaching (JiTT) is a Web-based teaching strategy that prepares both the student and the teacher for a more meaningful and engaging classroom encounter. We used this approach to teach statistics. When compared to an equivalent class based on content and textbook, the students in the JiTT class performed better on the final exam and expressed satisfaction with the approach. This study suggested the success of combining principles of active learning and Web-based technology.

Spence (2001) suggested that people learn best individually or in one-on-one relationships in which teacher and student are in constant communication with each other. The typical lecture course provides the teacher-to-student information flow, but the student-to-teacher information flow is minimal. Novak, Patterson, Gavrin, and Christian (1999) proposed a new teaching approach referred to as the Just-in-Time Teaching (JiTT) method. The method allows...
students to tell the teacher what they know just in time for class. The teacher uses the information to make the class period more engaging and instructive for the student.

Novak et al. (1999) developed the JiTT method to combine active learning and Web-based teaching techniques to teach physics. Two components of the JiTT method are essential. The first component asks instructors to post several short-answer or multiple-choice questions on the Web to probe student knowledge before they come to lecture. The second component asks instructors to read and use the responses in preparation for the class. The instructors can display the responses to the class to focus the lecture or class discussion. This approach facilitates an interactive learning environment and makes the class time more useful for both student and teacher. McKeachie (2002) argued that learning occurs when students see a relationship between what they are hearing from the instructor and what they already know. Research has shown the JiTT approach to increase learning and improve student attitudes toward the content of physics (Mzumara, Gavrin, & Chisholm, 2001; Novak & Patterson, 1998).

The JiTT approach shares characteristics with other effective teaching methods that aim to facilitate active learning. For instance, Conner-Greene (2000) made use of daily in-class essay quizzes to encourage regular reading of assignments as well as engagement with the material. Butler, Phillmann, and Smart (2001) sought to facilitate active learning for psychology students via the use of in-class writing exercises they collected about once a week. Comparatively, JiTT takes less class time and provides not only a summative evaluation but also an informative evaluation that instructors can use to create a more useful and meaningful classroom experience for students.

We applied the JiTT principles developed by Novak et al. (1999) to teaching statistics. We hypothesized that students who received the JiTT-based techniques would report greater involvement in the learning process and tend to learn more than students who did not receive the JiTT-based approach.

Method

Participants

Participants were students enrolled in two consecutive spring semesters of a sophomore-level psychological statistics course. Sixty-seven students in the control group took the course in the first spring semester. Fifty-six students in the JiTT experimental group took the course during the second spring semester.

Procedure

One of the authors taught both the JiTT and control groups at the 10:00 a.m. Monday/Wednesday/Friday time. This professor has taught this course for over 25 years. Both groups received the same course lectures, used the same textbook, participated in the same activities (with the exception of JiTT-specific activities), and took the same final exam.

The control group received a weekly 5-item in-class multiple-choice quiz on the textbook assignment prior to discussion of the material in class. The JiTT group did not have the weekly in-class quizzes. Instead, they responded to weekly (12) sets of preclass questions (PCQs) on an Internet site. An example of a PCQ set appears in Figure 1. Each PCQ set included two essay questions and one difficult multiple-choice question. Usually, two of the three questions required students to apply, analyze, or evaluate problems and procedures discussed in previous classes. The third question required students to complete assigned readings.

The instructor posted each PCQ set on the Web Wednesday afternoon and students had to submit their responses on a Web form by 8:00 a.m. Friday morning, 2 hr before their 10:00 a.m. class. During these 2 hr, the instructor read the responses and chose six or seven to anonymously display and discuss in class. The instructor made positive comments about each displayed response and, through discussion, asked the class to help remove possible misconceptions and to improve the answers. At the end of a discussion, the instructor presented the outstanding submitted answers to repeat the important ideas. The presentation of these outstanding answers gave students the opportunity to read examples of excellent writing by their peers. By Monday, each student received a graded copy of his or her submission.

We used both knowledge and attitude measures to judge the efficacy of the JiTT approach. The knowledge measure was the same cumulative final exam used at the end of both semesters and used problems that stimulated higher cognitive processing such as analysis and application. None of the problems had appeared in an earlier PCQ or in-class quiz. The affective
or attitude measure was composed of Likert-type questions in the end-of-semester student evaluation form.

Results

The average number of PCQ assignments submitted by the JiTT group was 11.24 of 12 (SD = 1.45), with a minimum value of 7 and a maximum value of 12. At the end of the semester, we asked the students in the JiTT group to evaluate their perceptions of the efficacy of the PCQs for learning and their enjoyment working with the approach. They responded on 5-point Likert rating scales. Table 1 shows the percentage of students who agreed or strongly agreed with each of the questions. In general, they were very positive about the use of this approach, particularly about discussing student answers in class. When we compared the final exam scores (possible range from 0 to 100), the JiTT class (M = 76.25, SD = 11.07) performed better than students in the control class (M = 72.39, SD = 8.89), t(119) = 2.13, p = .04, two-tailed, d = .38.

Discussion

The results supported our hypothesis that the JiTT students would like the approach and would find it facilitated their learning of statistics. Perhaps they were motivated to submit their best work because of the points they received and of the possibility that the instructor might show their responses to the class.

The higher final exam score for the JiTT group provided some empirical support for our second hypothesis. However, given the quasi-experimental nature of the design and the potential problem of experimenter bias, these results should be considered as suggestive rather than definitive.

Several drawbacks occur with the JiTT approach. First, instructors need to learn how to create Web forms and have a server to store them. Second, the approach takes some additional time to implement. This instructor spent an extra 1 to 2 hr per week preparing for class. Third, because these assignments were not due in class, some students complained about forgetting or missing the Internet submission deadline of 8:00 a.m. on Friday. Lastly, some students did not have convenient access to the Web.

The JiTT approach seemed to facilitate the teaching of statistics. We have also used it successfully in general psychology and research methodology classes. In general, students like using this approach in each of these classes. There is a question of whether this approach might be more effective in lower division or upper division courses that needs to be investigated. In conclusion, JiTT created a classroom experience that was meaningful, timely, and effective for both the instructor and the student. It provided immediate feedback to students about their level of understanding, and students liked using the method.

References


Note

Send correspondence to James O. Benedict, Department of Psychology, James Madison University, Harrisonburg, VA 22807; e-mail: benedijo@jmu.edu.

Table 1. Percentage Agreed or Strongly Agreed to Questions That Evaluated the JiTT Method

<table>
<thead>
<tr>
<th>Questions</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCQs were extremely effective for learning statistics</td>
<td>69</td>
</tr>
<tr>
<td>It was helpful for the professor to discuss answers in class</td>
<td>90</td>
</tr>
<tr>
<td>PCQs facilitated my understanding of statistical concepts</td>
<td>71</td>
</tr>
<tr>
<td>I enjoyed the PCQs because they were on the Web</td>
<td>56</td>
</tr>
<tr>
<td>I liked the PCQs because I could do them on my own time</td>
<td>67</td>
</tr>
<tr>
<td>PCQs encouraged me to read ahead</td>
<td>48</td>
</tr>
<tr>
<td>PCQs enhanced communication between student and professor</td>
<td>50</td>
</tr>
<tr>
<td>PCQs should be used in future statistics classes</td>
<td>77</td>
</tr>
</tbody>
</table>

Note. N = 55. We removed one student from the JiTT group because this student completed only 4 of the 12 PCQs during the semester. We also removed a student matched on the basis of the final exam score from the control group. JiTT = Just-in-Time Teaching; PCQ = preclass questions.

Using a “New Classic” Film to Teach About Stereotyping and Prejudice

Andrew N. Christopher and Jamie L. Walter
Albion College

Pam Marek
Anderson College

Cynthia S. Koenig
St. Mary’s College of Maryland

We describe a method for helping students learn about stereotype formation and prejudice by having them watch and discuss characters and scenes in the movie The Breakfast Club (Tanen &
Graduate admissions committees carefully examine transcripts; typically by a minimum of two faculty members.

Graduate admissions committees place a high value on transcripts, and either a low GPA or low GRE score may prompt a closer examination of a transcript.

One withdrawal does not appear to be a problem. Two withdrawals is probably not a problem, except for a minority of schools. For some institutions, withdrawals in particular courses are more detrimental than withdrawals in other courses.

After reading the respondents’ open-ended responses, however, it became evident that transcript evaluation is a complicated issue. Faculty examine the patterns of Ws over time, and it might make a difference if there are four Ws in one semester or one W in four consecutive semesters in the same course. Perhaps the type of class also interacts with the effect of a withdrawal—a general education course, psychology requirement, or a upper division elective. Stellar GRE scores or an exceeding high GPA may help to ameliorate the effects of Ws on transcripts. Future researchers interested in this topic might want to capture this complexity in an effort to help explain the impact of transcripts and withdrawals on the graduate admissions process.

Conclusions

Based on this research, what should I tell students who ask about the potential effect of withdrawals on graduate school aspirations? The following suggestions emerge from this study:

- Graduate admissions committees consider transcripts; typically by a minimum of two faculty members.
- Graduate admissions committees place a high value on transcripts, and either a low GPA or low GRE score may prompt a closer examination of a transcript.
- One withdrawal does not appear to be a problem. Two withdrawals is probably not a problem, except for a minority of schools. For some institutions, withdrawals in particular courses are more detrimental than withdrawals in other courses.

Comparing Bayes’s Theorem to Frequency-Based Approaches to Teaching Bayesian Reasoning

John Ruscio
Elizabethtown College

Despite the conceptual simplicity of Bayesian reasoning, people often err when calculating or estimating conditional probability. These mistakes can have significant real-world consequences, and
Bayes’s Theorem is a notoriously difficult remedy to teach. Experimen-
ters taught 113 students to use either Bayes’s Theorem, 2 × 2
tables, frequency grids, or frequency trees to solve a sample
mammogram problem. Immediately following written instruction,
group demonstration, and a question-and-answer session, perfor-
mance on new problems was equivalent across groups. However,
when retested 4 weeks later, participants in the Bayes’s Theorem
group solved fewer problems and demonstrated a poorer under-
standing of Bayesian reasoning than participants in all other
groups. Teaching a frequency-based approach to conditional prob-
ability appears to promote learning more effectively than teaching
Bayes’s Theorem.

Performing Bayesian reasoning can be essential to reach
sound conclusions, but most people—including many profes-
sionals (Abernathy & Hamm, 1995; Dowie & Elstein, 1988;
Eddy, 1982)—are not particularly good at it. Within psychol-
ogy, Bayes’s Theorem is often taught in graduate-level statist-
cs courses but seldom at the undergraduate level, especially
outside of introductory statistics courses. The underlying
principles are not complex, and teaching Bayesian reasoning
can allow instructors to show a large number of students how
psychological science can improve real-world decision mak-
ing through the development of mathematical aids to over-
come cognitive limitations and biases.

Although performing Bayesian reasoning is not demanding,
effectively teaching it remains challenging. Whereas students
trained to use Bayes’s Theorem fare poorly, people can learn
and implement techniques based on alternative formulations.
Recently, researchers have argued that humans evolved the
capacity to reason better with frequencies, which stem directly
from experience with the natural world, than with probabil-
ities, which are a relatively recent and more abstract human in-
vention (Cosmides & Tooby, 1996; Gigerenzer & Hoffrage,
1995). Preliminary data appear to support this assertion
(Sedlmeier & Gigerenzer, 2001).

For example, to determine the probability that a woman
with positive mammogram results has breast cancer, one can
construct a “frequency tree” (see Figure 1) that breaks down
a hypothetical sample of women first by the presence or ab-
sence of breast cancer and then by test results. It is then sim-
ple to establish how well mammography predicts cancer (e.g.,
8 out of a total of 107 women with positive tests had cancer,
for a probability of .075). A “frequency grid” is conceptually
similar to the frequency tree (see Figure 1), as is a 2 × 2 table
of test results by criterion status (Ruscio, 2002). These fre-
cuency-based techniques may be more meaningful and
computationally simpler than Bayes’ Theorem. The logic of
breaking down samples into subgroups using the base rate
and test validity is simple to grasp. The process itself draws at-
tention to the number of true and false positive cases, which
renders calculations trivial. Because Bayes’s Theorem inte-
grates probabilities in a way that is neither transparent nor in-
tuitive, it can be difficult to apply and there is little to be
gleaned from an incomplete attempt.

Sedlmeier and Gigerenzer (2001) achieved impressive re-
results with frequency trees and grids, but it is unclear how
readily their intensive training can be implemented. This ex-
periment tested the efficacy of four techniques—Bayes’s
Theorem, 2 × 2 table, frequency tree, and frequency
grid—in more ecologically valid ways.

Method

In exchange for course credit, 113 General Psychology stu-
dents took part in this experiment.¹ Two trained experiment-
ners conducted 14 sessions that ranged from n = 4 to n = 12.
Both experimenters conducted one or two sessions within
each experimental condition.

Prior to instruction, participants completed a pretest prob-
lem that contained a base rate (1% of women undergoing
mammography have breast cancer), a true positive rate (80%
of women with breast cancer test positive on the
mammogram), and a false positive rate (10% of women with-
out breast cancer test positive on the mammogram) and
posed a question requiring the calculation of positive predic-
tive power (the probability that a woman who tests positive
on the mammogram has breast cancer).

¹Demographic data were not collected. Given the nature of the
student body at Elizabethtown College and typical enrollment patt
After completing the pretest, participants read a single-page explanation of one method for solving Bayesian problems that illustrated the technique by correctly solving the pretest mammography problem. In step-by-step form, participants read how to locate the relevant information and apply either Bayes’s Theorem or a frequency-based method. Once all participants had read the instructions, the experimenter demonstrated the technique by working through the same steps outlined in the instructions on a large dry-erase board and answered any questions. When all participants felt prepared to work on new problems of this type, they took the first test, which contained three new Bayesian problems of the same form as the pretest problem. Most participants completed the first session in a total of 30 to 45 min.

Approximately 4 weeks later, 106 participants (94%) returned to take a second test with three new problems. Finally, participants completed a questionnaire that contained one additional problem to be solved without using a calculator or writing down any calculations and three questions that probed for a true understanding of Bayesian reasoning, as opposed to the mere ability to apply a mechanical technique that successfully solves problems. Participants completed the second session in about 15 to 20 min.

Results

In keeping with previous studies, “correct” responses had to lie within 5 percentage points of the true probability. Ten participants correctly solved the pretest problem; all subsequent analyses excluded these participants’ data, which were evenly distributed across experimental conditions.

The primary analysis tested for performance differences across experimental conditions via a 4 (experimental conditions: Bayes’s Theorem, 2 × 2 table, frequency tree, and frequency grid) × 2 (time: first vs. second session) mixed-model ANOVA. Scores on each test ranged from 0 (no problems solved correctly) to 1 (all three problems solved correctly).² There was no difference across conditions, $F(3, 92) = 1.42, p = .24, \eta^2 = .04$, but a large drop in performance over time, $F(1, 92) = 22.50, p < .01, \eta^2 = .20$. More important, whereas performance on the first test was comparable across instructional methods, performance deteriorated more markedly for those taught Bayes’s Theorem than for those taught a frequency-based technique, $F(3, 92) = 4.41, p < .01, \eta^2 = .13$ (see Figure 2).

To test the hypothesis that participants would learn and retain the frequency-based techniques better than Bayes’s Theorem, planned contrasts were conducted on performance at each time. The first contrast compared the Bayes’s Theorem group to all other groups; the second compared the 2 × 2 table group to the frequency grid and frequency tree groups; the third compared the frequency tree group to the frequency grid group. On the first test, none of these contrasts was significant, $t(99) < 1.65, p > .10, \eta^2 < .03$ for each. On the second test, participants in the Bayes’s Theorem group ($M = .06, SD = .13$) scored significantly worse than did those in the three frequency-based groups ($M = .32, SD = .39, t(92) = 3.47, p < .01, \eta^2 = .12$). Neither of the other two contrasts was significant, $t(92) < .59, p > .56, \eta^2 < .01$ for both.

A final analysis evaluated scores on the final questionnaire using the same three planned contrasts.³ Again, participants in the Bayes’s Theorem group ($M = .46, SD = .22$) performed significantly worse than did those in the three frequency-based groups ($M = .58, SD = .25, t(90) = 2.06, p = .04, \eta^2 = .05$, and neither of the other two contrasts was significant, $t(90) < .99, p > .32, \eta^2 < .02$ for both.

Discussion

The performance of participants taught one of the three frequency-based methods dropped only from 45% immediately following instruction to 32% 4 weeks later, whereas the performance of participants taught Bayes’s Theorem plummeted from 49% to 6%. Questions testing participants’ deeper understanding of the underlying logic of Bayesian reasoning revealed similar results. Neither performance on Bayesian problems nor an understanding of the logic of Bayesian reasoning differed significantly across the three frequency-based methods.

Whereas this pattern of results mirrors that obtained by Sedlmeier and Gigerenzer (2001), the absolute level of performance was considerably lower here. For example, in Sedlmeier and Gigerenzer’s first experiment, participants in the Bayes’s Theorem group solved about 60% of problems correctly immediately following training, and this figure dropped to about 20% over 5 weeks; in contrast, participants taught the frequency tree and grid techniques solved between 75% and 90% of their problems correctly, and performance did not drop over time. There are many possible reasons for the achievement gap across studies. Instruction time was briefer in this study (approximately 20 min) than in theirs (up to 2 hr). As compared to the written instructions, group demonstration, and question-and-answer session

²Scores on the first and second tests were scaled from 0 to 1 by dividing the total number of correct responses by 3 so that group means would correspond to the proportion of problems correctly solved. Scores on the final questionnaire, which ranged from 0 to 4, were divided by 4 so that group means would correspond to the proportion of correct responses.

³Two participants’ data were missing on the final questionnaire, so df for analyses of the final problem and follow-up analyses reflect slightly lower Ns than in previous analyses.

Figure 2. Mean test scores for both sessions and all four experimental conditions. Error bars represent +1 SE of the mean.
1. I thank the two experimenters, Jennifer Mills and Donna Ondik, for their invaluable assistance in collecting, entering, and checking these data. I am also grateful to Peter Sedlmeier for providing a large set of Bayesian problems from which the ones in this experiment were selected.

2. Send correspondence and requests for the full text of all experimental materials to John Ruscio, Department of Psychology, Elizabethtown College, One Alpha Drive, Elizabethtown, PA 17022; e-mail: rusciojp@etown.edu.

---

Teaching Successful Grant Writing to Psychology Graduate Students

Thomas Eissenberg
Department of Psychology and Institute for Drug and Alcohol Studies
Virginia Commonwealth University

Grant writing is a part of many psychology careers, but psychology departments seldom offer grant-writing courses. In 2000, Virginia Commonwealth University’s (VCU’s) Department of Psychology began offering a course focusing on the National Institutes of Health (NIH) National Research Service Award (F31). In 2 years, 16 students completed the course, and 6 submitted F31 proposals. Mean VCU priority scores were significantly better than the mean for all competing F31s submitted to the same institutes. NIH funded all 6 VCU proposals. This article describes the course and its key aspects. Grant-writing courses offer short-term rewards and potential long-term benefits to students, faculty, and psychology departments.

Grant writing is important for psychological research, researchers, and clinicians. For example, in 2001, total federal grant funding for psychological research was approximately $751 million (National Science Foundation, 2001). Psychologists’ success at writing grants can influence hiring, tenure, and promotion decisions, but the online catalogs of the top 24 psychology graduate programs (as ranked by U.S. News & World Report, 2002) showed no courses devoted to this topic. Other areas give more attention to training in grant writing. For example, the online catalog of Saint Louis University’s Department of Biochemistry and Molecular Biology reveals that PhD students are required to take a grant-writing course, and such a course is available to psychiatry postdoctoral fellows at the University of Pittsburgh (Reynolds et al., 1998). Thus, although psychology graduate students receive little grant writing training, students in other areas learn these necessary skills.

A formal course that teaches psychology graduate students how to write grants may be a vital but missing part of current curricula. A formal course might encourage students to apply for funding that will aid their graduate and professional careers, while increasing funding for psychological research in general. It would also allow instructors to receive acknowledgment for their pedagogic work and contribute to a higher and more standardized level of instruction.

This article describes a grant-writing course developed for graduate students in Virginia Commonwealth University’s (VCU’s) Department of Psychology. The course has two main goals: (a) teaching students how to write effective grant proposals and (b) increasing the amount of external funding for VCU’s graduate students.

References


Notes

1. I thank the two experimenters, Jennifer Mills and Donna Ondik, for their invaluable assistance in collecting, entering, and checking these data. I am also grateful to Peter Sedlmeier for providing a large set of Bayesian problems from which the ones in this experiment were selected.

2. Send correspondence and requests for the full text of all experimental materials to John Ruscio, Department of Psychology, Elizabethtown College, One Alpha Drive, Elizabethtown, PA 17022; e-mail: rusciojp@etown.edu.

Teaching of Psychology
Discussion

The results of this study, as presented in Table 1, provide an interesting look at what statistic instructors consider important. This Top 100 list may be construed as a first approximation of the core of introductory statistics in psychology. These results may also help instructors decide what the important terms to cover in class are and what terms are secondary. The approach of this study is unique because only undergraduate statistics instructors rated the importance of terms. Departments that offer multiple sections of statistics taught by different instructors using different textbooks might find these results useful in an effort to coordinate consistent coverage of the various content areas of statistics. The results of this study are more specific to psychology instruction than those from a previous study (Giesbrecht et al., 1997).

For some semesters, instructors may not cover certain chapters of information they consider less important; other semesters they may cover those chapters. What topics sometimes get deleted due to time constraints? The listing presented in Table 1 should help instructors and authors to answer these questions by understanding the core terms in the context of relative importance.

This study has limitations with regard to the methodological approach used, including the convenience sample of textbooks. In addition, in the instances where terms were highly similar but not identical, I included all terms. More work in this area is warranted to continue to verify the core terms. The listing provided here can be valuable for instructors in designing and making choices about the coverage of topics in an introductory course in statistics.

References


Notes

1. The comprehensive list of means and standard deviations for all 374 terms is available from the author.
2. Send correspondence to Eric Landrum, Department of Psychology, Boise State University, 1910 University Drive, Boise, ID 83725–1715; e-mail: elandru@boisestate.edu.

Promoting Conceptual Understanding of Statistics: Definitional Versus Computational Formulas

Katarina Guttmannova, Alan L. Shields, and John C. Caruso
University of Montana

Computer applications have replaced hand calculations as the relevant procedural skill for most of the statistical techniques in introductory statistics courses. Therefore, definitional formulas should replace computational formulas because only the former contribute to conceptual understanding. A review of 12 introductory statistics textbooks indicated that they emphasized computational formulas, particularly for complex techniques and exercises. We argue that the presentation and use of computational formulas is counterproductive to the learning goals of statistics courses and provide recommendations for instructors to facilitate the use of definitional formulas.

In the precomputer era, when students performed calculations by hand, computational formulas provided heuristics for difficult computations but sacrificed understanding of the underlying statistical concepts. Currently, however, 69% of psychology departments use computers in introductory statistics courses, and 90% use them for data analysis at some point in their curriculum (Bartz & Sabolik, 2001). Computers perform most statistical analyses. Thus we question the continued need for computational formulas and advocate instead for definitional formulas computed using hand calculators.

Although hand calculations are less tedious when using computational formulas, definitional formulas provide students with some perspective as to what the statistic is, when to use it, and how to interpret it. Consider, for example, the definitional,

$$s^2 = \frac{\sum (X_i - M)^2}{N},$$

and computational,
formulas for the variance, where \( X_i \) are the raw scores, \( N \) is the sample size, and \( M \) is the arithmetic mean of all observations. The variance of a distribution is a measure of how dispersed the scores are around the mean in the distribution. Hence, conceptual understanding is advanced by examining and using the definitional formula, whereas the computational formula is only a computational tool with no interpretative significance.

As Abelson (1995) observed, "From long observation of student struggles with statistics, I conclude that the difficulties lie not so much with computational mechanics as with an overall perspective on what they are doing" (p. xii). If instruction should focus on the tasks students will perform after completing their statistics course, and if computer applications have replaced hand calculations as the required procedural skill, then the advantage of the computational formulas is diminished. However, we hypothesized that a focus on computational in contrast to definitional formulas exists in current statistics textbooks, impeding students' grasp of the overall perspective.

Method and Results

We reviewed a convenience sample of 12 introductory statistics books published between 1999 and 2001 (see the Appendix). For each textbook, we tabulated the definitional and computational formulas in the text; in examples; and in exercises for chapters on variance, correlation, and one-way ANOVA. The results of our review appear in Table 1.

Among the textbooks we examined, there was not a statistically significant difference in preference for the definitional formula over the computational formula for the variance. Only 75% of the textbooks provided even a single exercise requiring the use of the definitional formula for the variance. We question how a student can conceptually understand the variance without ever using its definitional formula. In correlation chapters, authors of only 41.7% of the textbooks suggested that students work at least one exercise using the definitional formula. For one-way ANOVA chapters, only 33.3% of textbooks provided a numeric example, and only 25% provided even a single exercise using the definitional approach. We also performed an analysis of 26 textbooks published from 1990 to 2002 and reached similar conclusions (results are available from the first author).

Statistical tests (two-tailed, \( \alpha = .05 \)) of the difference in frequencies between computational and definitional formulas revealed that more textbooks employed computational formulas for correlation exercises (83.3% vs. 41.7%; \( z = 2.11, p = .04 \)) and for ANOVA examples (75% vs. 33.3%; \( z = 2.51, p = .01 \)). Furthermore, examining the pattern of frequencies in Table 1 reveals a clear trend: The frequencies for definitional formulas dropped as the content became more complex in text (100% for variance vs. 58.3% for ANOVA), \( z = 2.51, p = .01 \); in examples (100% for variance vs. 33.3% for ANOVA), \( z = 3.46, p < .001 \); and in exercises (75% for variance vs. 25% for ANOVA), \( z = 2.45, p = .01 \), whereas those for the computational formulas were relatively stable (never extending below 66.7% or above 83.3%).

Discussion

Our results show a strong emphasis on the computational approach at the expense of the definitional one as procedures became more complex and conceptual understanding is presumably more elusive. Our results indicate that textbook authors are emphasizing a set of formulas designed for hand calculations while sacrificing conceptual understanding. It is unlikely that current students will ever choose to use the computational formulas unless, of course, their statistics instructors require these procedures. Definitional formulas convey the meaning of the statistic at hand, and we believe that repeatedly observing and performing calculations with definitional formulas is beneficial to the attainment of conceptual understanding. However, we recognize that using hand calculations with definitional formulas can be tedious if the examples are not chosen carefully. Consequently, we recommend using manageable instructional datasets.

Table 1. Frequency of Occurrence of Definitional and Computational Formulas in Text, Examples, and Exercises in Three Selected Textbook Chapters

<table>
<thead>
<tr>
<th>Textbook Chapter</th>
<th>Text</th>
<th>Examples</th>
<th>Exercises</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Definitional formula</td>
<td>12</td>
<td>100.0</td>
<td>12</td>
</tr>
<tr>
<td>Computational formula</td>
<td>10</td>
<td>83.3</td>
<td>10</td>
</tr>
<tr>
<td>Correlation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Definitional formula</td>
<td>10</td>
<td>83.3</td>
<td>6</td>
</tr>
<tr>
<td>Computational formula</td>
<td>10</td>
<td>83.3</td>
<td>10</td>
</tr>
<tr>
<td>One-way ANOVA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Definitional formula</td>
<td>7</td>
<td>58.3</td>
<td>4</td>
</tr>
<tr>
<td>Computational formula</td>
<td>9</td>
<td>75.0</td>
<td>9</td>
</tr>
</tbody>
</table>

Note. \( N = 12 \).

*Indicates that the proportion (frequency/12) for definitional formulas is statistically different from the proportion for computational formulas (two-tailed test, \( \alpha = .05 \)).
Recommendations

The following recommendations to instructors of introductory statistics can make the use of definitional formulas easier for students:

1. Use small data sets: 5 or 10 observations are adequate to demonstrate most statistical techniques.
2. Present data in computationally manageable units: Instead of using a number such as 1.5 hr, use 90 min.
3. Construct data sets that produce integer means and standard deviations: Dudek (1996), McGown and Spencer (1996), and Read and Riley (1983) provided methods for generating such data sets.
4. Use a textbook that emphasizes the definitional approach: Aron and Aron (1994) professed the perspective most similar to what we are proposing. They emphasized definitional formulas throughout and stated that: “The purpose of this book is to help you understand statistical procedures, not to turn you into a computer by having you memorize and use computational formulas you will rarely, if ever, employ again” (p. 47). Abrami, Cholmsky, and Gordon (2001) also stressed the definitional approach. Those who would like to use Web-based instructional methods and materials should consider Stockburger (1998).

Conclusions

Historically, using both definitional and computational formulas for teaching statistical concepts in introductory statistics classes was justified. The calculation of statistics with pencil and paper or calculator was required of students after the course; hence learning computational formulas was a relevant objective. However, given the widespread use of computers for instructional and especially noninstructional data analysis, the computational formulas have lost their primary purpose. To help students gain conceptual understanding, we believe that instruction as well as statistics textbooks’ text, examples, and exercises should be designed around procedures developed specifically for that purpose: definitional formulas.

References

   Data sets with integer means and standard deviations. In M. E. Ware & D. E. Johnson (Eds.), Handbook of demonstrations and activities in the teaching of psychology (Vol. 1, p. 87). Mahwah, NJ: Lawrence Erlbaum Associates, Inc. (Originally published in Teaching of Psychology, 7, 63)

Appendix

Introductory Statistics Textbooks Reviewed


Notes

1. Alan L. Shields is currently an Instructor of Psychology in the Harvard Medical School and works in the Alcohol and Drug Abuse Treatment Program at McLean Hospital. John C. Caruso is currently the Director of Services at Dissertation Statistical Services. Katarina Guttmannova is currently a postdoctoral fellow at the Institute for Policy Research, Northwestern University.
2. Send correspondence to Katarina Guttmannova, Institute for Policy Research, Northwestern University, 2040 Sheridan Road, Evanston, IL 60208–4100; e-mail: k-guttmannova@northwestern.edu.

Graphing Psychology: An Analysis of Source Material of Graphs in Introductory Psychology Textbooks

Jeffrey S. Nevid and Nicholas Forlenza
St. John’s University

We conducted an analysis of source material in data graphs included in introductory psychology texts. Our findings showed an increasing trend toward inclusion of graphs representing more re-
author confronted a case of academic dishonesty by one of her students, the faculty successfully guided her through the challenging university hearing and appeals process (Keith-Spiegel, Wittrig, Perkins, Balogh, & Whitely, 1993).

The students who have taken these GTA pilot courses have provided valuable feedback in the form of midterm and postcourse evaluations. Students consistently report they appreciate the attention they receive in smaller “seminar-style” classes of 35 to 60 students in contrast to the 100- to 300-student lecture courses. They particularly value the opportunity to have more specialized psychology courses taught by knowledgeable and enthusiastic instructors. At the time of publication, seven sections of one particular special topics course enrolled 450 students over five semesters. Themes evident in course evaluations included appreciation for enthusiasm and accessibility of GTAs and for specialized knowledge about their current areas of study. These courses thus appear to provide students with some of the advantages of the smaller classes typically found at smaller, liberal arts colleges. At the same time, students have access to the additional resources of the large research-oriented institutions.

In summary, these courses meet a need of the psychology department to offer additional upper division electives to both majors and nonmajors. Budget restraints can limit the ability to offer these valuable courses without the assistance of willing doctoral students. Graduate students who develop courses also benefit from the opportunity to hone a new skill set. We can now offer our graduate students more steps on a developmental track toward becoming outstanding educators. This type of mutually beneficial relationship is an example of how university departments can thrive in economically difficult times while supporting their junior colleagues in becoming more skilled teachers of psychology.

References


Notes

1. We thank the anonymous reviewers for their thoughtful comments.

2. Send correspondence to Kirsten T. Michels, Department of Psychology, University of Georgia, Athens, GA 30605; e-mail: kirstemichels@uga.edu.

Avoiding Confusion Surrounding the Phrase “Correlation Does Not Imply Causation”

Julie Hatfield
New South Wales Injury Risk Management Research Centre
University of New South Wales

Gavin J. Faunce and R. F. Soames Job
University of Sydney

Psychology textbooks, teachers, and students frequently employ phrases like “correlation does not imply causation.” Because of the statistical meaning of the term correlation, students often misunderstand such phrases to mean that limitations on causal inference arise from the type of statistical analysis conducted. Our suggestions for teaching about the causal implications of different research designs aim to avoid confusion that is likely to be detrimental to psychology.

Psychology textbooks and teachers frequently offer students the lesson that “correlation does not imply causation,” perhaps because it is simple and easily remembered. For example, in their introductory psychology textbook Gleitman, Fridlund, and Reisberg (1999) stated that “correlation does not imply causation” (p. A13) and that “correlations cannot by themselves signal a cause and effect relationship” (p. A13). In an earlier edition of the same textbook, Gleitman (1991) stated that “correlations can never provide solid evidence of a causal link” (p. A19). In another introductory psychology textbook, Wade and Tavris (1998) claimed that “correlation does not show causation” (p. 57). More recently, Smith, Bem, and Nolen-Hoeksema (2001) stated that “when two sets of scores are correlated, we may suspect that they have some causal factors in common, but we cannot conclude that one of them causes the other” (p. 530).

The Logical Intended Meaning of “Correlation Does Not Imply Causation”

When correlation is taken as a synonym of association, these claims have logical merit. An association between two variables, A and B, can mean any of three things (after dismissing the possibility of a Type I error via available statistical
techniques). First, changes in A may cause changes in B. Second, changes in B may cause changes in A. Third, changes in other, possibly unknown variables, say C, may cause changes in both A and B independently. Mere observation of an association between A and B does not allow discrimination among these three causal possibilities, so manipulation is necessary to allow inferences regarding causal direction. For example, the hypothesis that A causes B is best supported by a finding that unconfounded manipulation of A results in corresponding changes in B.

Thus, claims such as "correlation does not imply causation" identify a weakness of nonexperimental research designs relative to experimental research designs. In experimental research, the investigator deliberately manipulates at least one independent variable and observes the effects on at least one dependent variable. In nonexperimental research (unfortunately often referred to as correlational research), the investigator merely observes selected variables rather than manipulating any directly. It is this lack of manipulation that restricts causal inference.

Confusion Arising From the Statistical Meaning of Correlation and Clarification

Because correlation has a statistical meaning, psychology students are apt to misunderstand statements such as "correlation does not imply causation" as meaning that the limitations on causal inference are the fault of the statistic, rather than of the research design. We have often witnessed this confusion in students, but it is also apparent in textbooks. Although the textbooks we quoted earlier used the term correlation ambiguously, some statistics textbooks refer explicitly to the correlation statistic when making claims about the causal implications of correlation. For example, Welkowitz, Ewen, and Cohen (1976) stated that "you cannot determine the cause … from the correlation coefficient" (p. 159). Similarly, Kirk (1978) suggested that a "common error in interpreting a correlation coefficient is to infer that because two variables are correlated, one causes the other" (p. 108).

Limitations of Nonexperimental Research Are Not Avoided by Avoiding the Correlation Coefficient

Kirk (1978) was correct in identifying the common misinterpretation of correlation coefficients (in nonexperimental research). However, the phrasing tends to suggest that it is the fault of the statistic. It would be clearer to replace the word correlated with associated and to identify that other statistical techniques (e.g., regression, ANOVA, t test, chi-square, probit analysis, nonparametric techniques) also do not allow causal inferences from nonexperimental data. For example, the t test or one-way ANOVA is appropriate to test the relation between two observed (but not manipulated) variables when one is dichotomous and the other continuous and causal inferences are not possible. A mere observation that people who do crossword puzzles have lower cholesterol than people who do not do crossword puzzles (using a t test) does not imply that doing crossword puzzles reduces cholesterol. One plausible interpretation might be that people of higher socioeconomic status have both more time to do crossword puzzles and a better diet.

Furthermore, avoiding the correlation statistic in nonexperimental research does not miraculously remove the limitations on causal inferences that are inherent in such research. We have had students suggest that they dichotomize continuous variables so that they may make causal inferences from nonexperimental data by using a t test. Of course, this approach will not work. For example, findings that people with high residential noise exposure have lower sleep quality and psychosocial well-being than people with low residential noise exposure are causally ambiguous regardless of whether researchers detect this association using a t test (e.g., Öhrström, 1991) or using the full range of sound level measures and a correlation. People who live in high noise areas may be of low socioeconomic status, which may have an independent effect on their psychosocial well-being. Nonetheless, making a continuous variable dichotomous may be justified by a need to increase statistical power (Zimmerman & Zumbo, 1992).

The Correlation Coefficient Can Appropriately Lead to Causal Inferences in Some Research Designs

Welkowitz et al. (1976) were incorrect in suggesting that a correlation coefficient never allows determination of causal direction. In an experimental design in which a researcher manipulates A to investigate corresponding changes in B, a correlation coefficient supports an inference of causation as effectively as other statistical techniques. For example, in a repeated measures design, Berglund, Berglund, and Lindvall (1976) exposed participants to noise stimuli of several different sound levels and found a strong correlation between sound level and self-reported annoyance (r around .9). This finding provides compelling evidence that noise exposure influences annoyance. Of course, other statistical techniques (e.g., t tests or ANOVA) are more common than correlational analysis in experimental research because A is generally manipulated to only a few levels.

Furthermore, there are instances in which correlations (and other statistical techniques) may allow causal inferences from nonexperimental data. A researcher can appropriately infer that A causes B if she observes an association between A and B and can eliminate the alternative explanations (i.e., B causes A; C causes A and B) on other grounds. For example, the explanation that B causes A is less likely if A precedes B in time or if there is no logical mechanism to support this explanation. For instance, if A is gender and B is some attitudinal variable, it is illogical to argue that attitude causes gender (unless the attitude influences the gender composition of the sample).

Eliminating the explanation that C causes A and B independently is typically more difficult. This explanation may be undermined by demonstrating that all reasonable probabilities for C (which may not be simple to identify) are not associated with A and B. Even if C is associated with A and B, informed commentary is possible. For example, associations between negative attitudes toward aviation and annoyance with aircraft noise (see Job, 1988) may not indicate that such attitudes worsen annoyance but rather that noise exposure worsens both variables. Indeed, noise exposure is associated with both attitude toward aviation and annoyance with aircraft (Job, 1988). Nonetheless, Job et al. (1998) found that
negative attitudes toward aviation had an equivalent impact on annoyance regardless of noise exposure. These findings strongly suggest that the association between attitudes and annoyance is partially independent of noise exposure (although the causal direction between them remains unclear).

Suggestions for Avoiding Confusion

The intended meaning of the phrase “correlation does not imply causation” is important, and the confusion that is apt to result from it may be detrimental by contributing to inappropriate conclusions from research results. Thus, all teachers of psychology should take care to avoid the confusion surrounding the phrase by adopting the following recommendations:

1. Employ the term correlation only in a statistical context, and then refer specifically to correlational analysis or the correlation coefficient.
2. Employ the term association to describe the relation between two variables without identifying the particular statistic employed to detect the relation.
3. Refer to research designs in which researchers merely observed variables (rather than manipulating them) as nonexperimental designs rather than correlational designs.
4. Replace phrases like “correlation does not imply causation” with the phrase “without manipulation, association does not imply causation.” Given people’s predilection for catchy phrases, it appears wise to offer a suitable alternative to the flawed but commonly used existing phrase. Our alternative phrase correctly stresses the importance of the type of research design, rather than the type of statistical analysis performed. Nonetheless, a more substantial phrase such as “association without manipulation does not normally imply causation” would better account for circumstances in which causal inferences from nonexperimental data are defensible. A blander phrase, such as “nonexperimental designs do not imply causation” suffers from not being sufficiently catchy.

Conclusions

In our experience, many students mistakenly form the impression during their undergraduate years that limitations on causal inference are imposed by the correlation statistic rather than by nonexperimental research designs. Although objective data are not yet available, casual observation suggests that through application of the recommendations offered in this article, teachers of psychological statistics and research design can help to eliminate such errors and facilitate more appropriate inferences from nonexperimental and experimental research.

References


Notes

1. R. F. S. Job is currently General Manager of Road Safety Strategy Branch, Road and Traffic Authority of New South Wales, Australia.

2. Send correspondence to Julie Hatfield, NSW Injury Risk Management Research Centre, University of New South Wales, New South Wales, 2052, Australia; e-mail: J.Hatfield@unsw.edu.au.

An Active Learning Classroom Activity for the “Cocktail Party Phenomenon”

Michael A. Clump

Marymount University

This article presents an active learning demonstration of the “cocktail party phenomenon.” It involves dividing the students in the class into groups of 3, with 2 individuals acting as speakers and 1 person as the participant. By simultaneously involving all of the students, more students experience the effect in an environment that replicates the cocktail party phenomenon and the students’ experiences outside of class, such as in a coffee house. The 3 within-subjects conditions illustrate how certain information slips through an attentional block. The students’ evaluations indicated the technique was enjoyable, useful, and a good way to learn about the topic.

Everyone experiences a time when they “tune out” someone else while talking on the phone or watching television. In addition, everyone listens at a party to one person tell a story, but occasionally eavesdrops on another conversation. Cherry (1953) and Cherry and Taylor (1954) began the study of selective attention by using a laboratory task in which participants immediately repeated information presented to one ear, known as shadowing, and ignored information presented...
Improving Attitudes Toward Statistics in the First Class

Robert A. Bartsch
University of Houston–Clear Lake

I present an activity designed to improve students’ attitudes toward statistics. On the first day of class, students drew numbers from bags and then answered statistical questions about the numbers. I designed the questions to illustrate that students already intuitively knew how to estimate many different statistics. Evaluations from 94 students demonstrated that the activity improved their attitudes toward statistics.

Most students are apprehensive about statistics (Sciutto, 1995), and math anxiety and low self-efficacy correlate with poorer performance in math classes (e.g., Benson, 1989; Feinberg & Halperin, 1978; Ma, 1999; Pajares & Kranzler, 1995; Schutz, Drogosz, White, & Distefano, 1998). Methods to improve attitudes include showing students that statistics are easier or more interesting than expected. For example, Connor (2003) described how to use students’ bodies to graphically view data, Johnson (1989) demonstrated ANOVA on an intuitive level, Weaver (1999) related statistics to familiar objects, Zerbolio (1989) taught sampling distributions by having students visualize scores as marbles in a bag, and Sgoutas-Emch and Johnson (1998) showed that journal writing reduces math anxiety.

I propose another activity to improve attitudes toward statistics. Unlike other activities, this demonstration can cover a variety of different statistical concepts. Furthermore, an instructor can use this activity on the first day, which helps to set the right tone for the semester (Jacobs, 1980).

Activity

Students (71 women and 23 men) from three undergraduate (n = 54) and two master’s-level (n = 40) statistics classes from two small state universities participated in this activity. On the first day of class, I asked students to rate their current fear of statistics, their ability to do statistics, and their ability to understand statistics using a Likert scale ranging from 1 (very low) to 9 (very high). Ratings were highly related (α = .85), and analyses for each individual measure produced similar conclusions. Therefore, I reversed the fear score and computed the mean of the three ratings to produce a statistics attitude score.

I introduced the activity by explaining that students would draw a slip with numbers on it from a brown paper bag, and I would ask questions (see Table 1) about these numbers. Approximately 5 to 10 students drew slips and read their numbers aloud. I used the small n to move the demonstration along quickly. I wrote the numbers on the board or had the students write the numbers on their paper. I then asked the entire class questions about the selected numbers.

I first used slips containing single numbers to talk about the mean. Students usually gave reasonable estimations, and I briefly differentiated what reasonable and unreasonable estimations would be. For example, a reasonable estimation of the mean had to be between the highest and lowest number and should be near the middle of the data. I then asked students the approximate average difference between the numbers and the mean. After students provided a reasonable estimation, I informed them that they had just estimated a standard deviation and again described good and bad estimates. For example, a poor estimate of the standard deviation would be the difference between the mean and the most extreme number. Typically, I addressed these questions three times with different bags of numbers.

I then continued through the questions listed in Table 1. I demonstrated that with a larger sample size, a person has increased confidence in the population value of the mean and standard deviation of the numbers in a bag. I also modeled a one-sample t test by having students draw numbers and asking them whether all the numbers in the bag had an average of a particular value. In this case, as with other inferential statistics, students did not estimate the value of the statistic but rather attempted to determine the statistical conclusion.

For independent t tests, students drew a slip from two bags, and I asked if numbers from one bag were different from numbers in another bag. To demonstrate p values, I asked students the percentage chance that the average of the numbers in the two bags were the same.

Table 1. Individual Questions Asked During Activity

<table>
<thead>
<tr>
<th>Bags Used</th>
<th>No. Per Slip</th>
<th>Topic</th>
<th>Question Asked</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Mean</td>
<td>What is the approximate mean of the distribution?</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>Standard Deviation</td>
<td>What is the approximate distance between each number and the mean?</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>Large versus small n</td>
<td>(Asked after each number) What is the approximate mean and standard deviation?</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>One sample t test</td>
<td>Do the numbers in this bag differ from [pick number]?</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Independent t test</td>
<td>Which bag has the highest overall mean?</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>p values</td>
<td>What is the likelihood that the means of the two bags are the same?</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>Dependent t test</td>
<td>Which number is higher, the first or the second?</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>Correlation</td>
<td>As the number on the right increases, what happens to the number on the left?</td>
</tr>
</tbody>
</table>
| 1         | 2            | Regression             | What would you predict the number on the left to be if the number on the right was [pick number]?
The slips with two numbers represented scores on two variables from the same participant. I used these slips to demonstrate concepts such as dependent t tests (by asking if the first number was different from the second number), correlation (by asking if there was a relation between the two numbers), and regression (by predicting the second number given the first number).

I continually emphasized to the students that they had just demonstrated that they already intuitively knew how to estimate these statistics. At the conclusion of the activity, I went over the list of all statistics covered in the activity and stated how students could make an educated guess that demonstrated they intuitively understood the statistical concepts. To evaluate this activity, I readministered the statistics attitude scale. The entire activity took about 30 min to complete.

Evaluation

As an overall evaluation, I first performed a dependent t test examining the attitude scores before and after the exercise. Overall, students had more positive attitudes after the activity (pre M = 5.11, SD = 1.95; post M = 6.23, SD = 1.53), t(93) = −9.71, p < .001, d = 0.61.

I also investigated a possible limitation. Specifically, I believed that students with positive attitudes toward statistics would benefit less from this activity. I created an attitude improvement variable by subtracting the pretest score from the posttest. I then regressed the pretest scores onto attitude improvement, r(92) = −.62, p < .001, to produce the regression equation:

\[
\text{Predicted Attitude Improvement} = -0.36(\text{Pretest Score}) + 2.94
\]

This equation predicts that students who had negative attitudes toward statistics at the time the course begins would improve more than students who had positive attitudes.

Discussion

This 30-min activity improved attitudes toward statistics, especially for students who were uncomfortable with statistics. By conducting this activity at the beginning of the semester, students can start the semester less anxious.

Another key benefit is that instructors can easily modify this activity. They can use their own questions for the topics they believe are most important or conduct this activity any time they want to illustrate students’ intuitive understanding of certain concepts. For example, this activity could demonstrate the basic idea of confidence intervals or ANOVA. Furthermore, instructors can manipulate the numbers in the bags to demonstrate more complicated concepts such as how differences in the standard deviation affect the perceived likelihood of rejecting the null hypothesis.

This activity has some limitations. By itself, it is unlikely to produce lasting changes in students’ attitudes toward their statistical abilities (Sciutto, 1995). Also, this activity is most effective with students who have negative attitudes toward statistics. As indicated, students with positive attitudes receive less benefit. Nevertheless, this activity is beneficial for improving attitudes for students from a variety of different backgrounds in a variety of institutional settings from the beginning of the course.

References


Notes

1. I thank Jo Meier, Beth Hentges, and three anonymous reviewers for their comments on a previous version of this article.

2. Send correspondence to Robert A. Bartsch, University of Houston–Clear Lake, Box 15, 2700 Bay Area Boulevard, Houston, TX 77058–1098; e-mail: bartsch@uhcl.edu.
Retrieving Essential Material at the End of Lectures Improves Performance on Statistics Exams
Keith B. Lyle and Nicole A. Crawford
Teaching of Psychology 2011 38: 94
DOI: 10.1177/0098628311401587

The online version of this article can be found at:
http://top.sagepub.com/content/38/2/94
Retrieving Essential Material at the End of Lectures Improves Performance on Statistics Exams

Keith B. Lyle¹ and Nicole A. Crawford¹

Abstract
At the end of each lecture in a statistics for psychology course, students answered a small set of questions that required them to retrieve information from the same day’s lecture. These exercises constituted retrieval practice for lecture material subsequently tested on four exams throughout the course. This technique is called the PUREMEM (pronounced “pure mem”) procedure for Pure Memory or Practicing Unassisted Retrieval to Enhance Memory for Essential Material. Exam scores were significantly and substantially higher in a section of the course taught with the PUREMEM procedure than one taught without it. Students liked the procedure and believed it increased learning via several different mechanisms.

Keywords
learning, memory, retrieval, statistics, testing effect

Instructors spend many hours developing lectures, and students spend many hours attending them, but students’ retention of lecture material is seemingly often poor. Here we present a technique for enhancing retention of essential lecture material by requiring students to retrieve the material from memory before leaving class. We were inspired by research showing that retrieving information from memory soon after acquisition increases long-term retention (Roediger & Karpicke, 2006). For example, Butler and Roediger (2007) showed that taking a short-answer test covering lecture material immediately after lectures increased long-term memory for the material by 31% relative to studying a summary of the material and by 135% relative to doing nothing with the material. Taking a test soon after acquiring information is a form of practicing retrieval in preparation for future tests and retrieval attempts. Butler and Roediger’s results, like many similar ones, suggest that practicing retrieval may increase the retention of educationally relevant material in real classrooms (McDaniel, Roediger, & McDermott, 2007).

Research demonstrating the value of retrieval practice caused us, like others (e.g., Gates, 1917; Jones, 1923), to see tests as not only assessment devices but also potential learning aids for lecture material. Therefore, we incorporated retrieval practice into a course on statistics for psychology by administering brief tests requiring retrieval of essential lecture material at the end of each lecture. We dubbed the tests PUREMEMs (pronounced “pure mem”), which we told students they could think of as standing for either pure memory or, more formally, Practicing Unassisted Retrieval to Enhance Memory for Essential Material. We avoided the labels tests and quizzes because these words primarily connote assessment rather than memory enhancement.

The PUREMEM procedure’s explicit emphases on retrieval practice versus assessment and on memory for lecture material versus assigned readings differentiate it from other procedures involving frequent testing (cf. Landrum, 2007; Narloch, Garbin, & Turnage, 2006; Padilla-Walker, 2006). For example, in Leeming’s (2002) exam-a-day procedure, students take an exam covering material from the preceding lecture and a textbook reading at the start of every class. The exams have a critical assessment function because scores on them are central to students’ course grades.

Next we describe the PUREMEM procedure in detail and compare exam performance in two sections of a statistics course with and without PUREMEMs. We also report data on students’ attitudes toward PUREMEMs and beliefs about how PUREMEMs enhanced learning. Theoretically, PUREMEMs enhance learning (specifically, retention) by inducing retrieval practice, but having retrieval practice at the end of lectures may help in additional ways, such as by causing students to attend more during lectures. We had students evaluate several mechanisms by which PUREMEMs conceivably may affect learning.

¹ University of Louisville, Louisville, KY

Corresponding Author:
Keith B. Lyle, University of Louisville, Department of Psychology, 2301 S. 3rd St., Louisville, KY 40292
Email: keith.lyle@louisville.edu
Method
Participants
The first author taught two sections of an undergraduate course on statistics for psychology in consecutive years. Only the second section included PUREMEMs. The first section served as a comparison. There were initially 77 students in the PUREMEM section and 78 in the comparison section, with one withdrawal in the former (1.3%) and five in the latter (6.4%), a nonsignificant difference, $\chi^2(1) = 1.52, p = .22$. Students in the two sections were indistinguishable on demographic and academic variables collected anonymously on a university-mandated course evaluation form. Across the two sections, 70% were 20 to 24 years old, 75% were female, 19% were minority or international students, 48% had at least a 3.0 GPA, 73% had accumulated at least 51 credit hours, and 41% had employment or other extracurricular commitments for 21 to 40 hours per week (for tests of between-section differences, all $ps > .15$).

Procedure
Class format. The courses had two 75-minute lectures per week. The style and content of lectures were the same in the PUREMEM and comparison sections, as were the syllabus and grade scale. Students took four noncumulative, multiple-choice exams covering the lecture material; exams were evenly distributed throughout the semester. Exams in the two sections covered the same material, were of comparable length, contained four-alternative questions, and consisted of a mixture of computational, noncomputational, and application problems. Between sections, there were slight differences in exams restricted to superficial elements of the problems, such as specific numbers in computational problems or details of scenarios in application problems.

 PUREMEM procedure. Administration of PUREMEMs occurred after each lecture during the final 5 to 10 minutes of class. Before the PUREMEM, students put away notes and took out paper on which to write answers. Then several questions—most often four, but ranging from two to six—were projected from a PowerPoint slide (see Table 1 for examples). The questions probed material from the same day’s lecture only, specifically material the instructor deemed most essential. For most questions, the correct answer consisted of one or two words, a number, or a simple figure. We did not use multiple-choice questions because they may not enhance long-term retention as much as do more open-ended questions (McDaniel et al., 2007). Students submitted their answers before leaving class. We projected the correct answers from a PowerPoint slide at the start of the next class and posted them on a course website. Students could review the answers and ask questions before the new lecture. If there were no questions, the feedback phase did not consume any class time.

Although assessment was not the primary purpose of PUREMEMs, the second author scored them, and they had a small weight in the grading scheme to motivate students to take them seriously. Students’ mean score for all PUREMEMs counted for 8% of their overall course grade. There were 21 PUREMEMs in total, so individual PUREMEMs counted for only .38% of one’s overall grade. Because individual scores had little impact on grades, we did not allow students to drop or make up any low or missing scores, thereby minimizing administrative load. Students received a zero for missed PUREMEMs.

At the end of the semester, students in the PUREMEM section anonymously completed an evaluation form concerning PUREMEMs. Table 2 shows the questions on the form. Sixty-five students (86%) completed the evaluation.

Table 1. Example PUREMEMs Related to Each of the Four Exams

<table>
<thead>
<tr>
<th>Exam</th>
<th>PUREMEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>All distributions of variables show two things: The values the variables could possibly have and what else about the values?</td>
</tr>
<tr>
<td>2</td>
<td>Frequency tables can be made to organize data from which type of variable(s)?</td>
</tr>
<tr>
<td>3</td>
<td>Histograms can be made to organize data from which type of variable(s)?</td>
</tr>
<tr>
<td>4</td>
<td>Without worrying about any details, draw the basic shape of a unimodal, symmetrical distribution.</td>
</tr>
</tbody>
</table>

Lyle and Crawford

95
Table 2. Mean Student Evaluations of PUREMEMs (with SDs)

<table>
<thead>
<tr>
<th>Question</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>At the present time, how would you describe your overall feelings about the use of PUREMEMs in this course?</td>
<td>5.7 (1.4)</td>
</tr>
<tr>
<td>To what extent do you think PUREMEMs helped you in this course specifically by...</td>
<td>6.0 (1.3)</td>
</tr>
<tr>
<td>Giving you a chance to practice answering questions that were similar to those on problem sets and exams?</td>
<td>5.9 (1.5)</td>
</tr>
<tr>
<td>Helping you identify some of the important topics that were going to be covered on problem sets and exams?</td>
<td>5.9 (1.5)</td>
</tr>
<tr>
<td>Causing you to come to lecture more than you otherwise would have?</td>
<td>5.8 (1.5)</td>
</tr>
<tr>
<td>Causing you to pay more attention during lecture than you otherwise would have?</td>
<td>5.8 (1.5)</td>
</tr>
<tr>
<td>Giving you a better sense of how much you had learned in each lecture than you otherwise would have had?</td>
<td>5.8 (1.3)</td>
</tr>
</tbody>
</table>

a. The response scale ranged from 1 (very negative) to 7 (very positive).
b. The response scale for the five questions below ranged from 1 (very little) to 7 (a lot).

Relationship between PUREMEM and exam questions.

PUREMEM questions targeted essential lecture material and, logically, exams included questions that targeted the same material. However, PUREMEM questions did not appear verbatim on exams. As stated, PUREMEM involved short answer questions, whereas exams had multiple-choice questions. PUREMEM questions were often the basis for the stems of exam questions, but although the conceptual focus was the same in those cases, the exact wording or numbers differed. For example, the exam question targeting the same material as the third example PUREMEM question in Table 1 was “Which of the following is a way to show the distribution of a numeric variable? A) A histogram, B) A bar graph, C) Both (A) and (B), D) Neither (A) nor (B).”

Results

Retrieval practice can increase the long-term retention of information but presumably only if retrieval attempts are successful (e.g., Kang, McDermott, & Roediger, 2007). To determine whether students could retrieve lecture material immediately after its presentation, we calculated each student’s mean PUREMEM score, excluding scores of 0 for missed PUREMEMs because there was no retrieval attempt on those exercises. The class mean was 86% (SD = 10%). Thus, retrieval attempts on PUREMEMs were usually successful.

To analyze exam performance, because proportions often violate the homogeneity and normality assumptions of inferential statistical tests, we applied the arcsine transformation to proportion correct on exams. Significance test statistics present analyses using the transformed proportion as the dependent measure. Descriptive statistics present raw proportions to retain interpretability. We submitted proportion correct on the four exams to a 2 (section) × 4 (exam) mixed-design ANOVA in which the first factor was between participants and the second was within. We included exam scores from only those students who took all four exams (nS = 74 and 70 in the PUREMEM and comparison sections, respectively). Overall, exam scores were significantly higher in the PUREMEM section (M = .86, SD = .09) than in the comparison section (M = .78, SD = .14), F(1, 142) = 11.23, MSE = .42, p = .001, ηp² = .073. There was also a significant main effect of exam, F(3, 426) = 8.65, MSE = .066, p < .001, ηp² = .057, and, more importantly, a significant interaction, F(1, 142) = 13.07, MSE = .066, p < .001, ηp² = .084. The interaction was such that the mean score on Exams 1, 2, and 4 was .08 to .13 higher in the PUREMEM section. The PUREMEM advantage was significant for each of these exams individually, smallest t(142) = 3.16, p = .002, Cohen’s d = 0.55. However, mean scores were identical in the two sections on Exam 3. Inspection of the means revealed that students in the comparison section scored markedly higher on Exam 3 than on any of the other exams, suggesting that students can sometimes achieve good retention of lecture material without end-of-lecture retrieval practice. Nevertheless, PUREMEMs substantially improved performance on the majority of exams. Correspondingly, far fewer students earned mean exam scores lower than 70% in the PUREMEM section (5.4%) than in the comparison section (27.1%).

Controlling for number of PUREMEMs taken, which indexes lecture attendance, the partial correlation between mean PUREMEM scores (excluding missed PUREMEMs) and mean exam scores was significantly positive, r(71) = .66, p < .001.

As shown in Table 2, students’ attitudes toward PUREMEMs were very positive, and students believed that PUREMEMs were of considerable help in all five ways queried.

Discussion

Adding a brief retrieval exercise for essential lecture material at the end of every lecture in a statistics course significantly and substantially increased exam scores. Students liked the...
retrieval practice and believed it was helpful. Specifically, students indicated that PUREMEMs helped them identify important topics, monitor their learning, increase their attendance, and pay more attention to lectures. Although we obtained these benefits in a statistics course, we see no reason not to expect that PUREMEMs would produce similar benefits in any course involving informationally rich lectures.

The PUREMEM procedure is potent and simple to implement, but concerns may exist about the amount of time necessary to administer and score PUREMEMs. Leeming (2002) addressed the same concerns regarding the exam-a-day procedure mentioned in the introduction, and our response is similar. First, it was possible to compensate for time spent in class on PUREMEMs by delivering more focused lectures. Although the 5 to 10 minutes devoted to PUREMEMs could be spent presenting additional information, we see no point in presenting information if the likelihood of retention is low. Second, because PUREMEMs were brief and it was not necessary to provide written feedback, scoring took only about 30 minutes.

We conclude by observing that retrieval practice is seemingly a highly flexible means of enhancing memory for lecture material. We used one particular method of implementing retrieval practice, but we can imagine numerous variations. Aspects of the procedure that different instructors might alter to better suit the needs of their own teaching styles or class formats include, for example, the number of questions, whether questioning occurs spaced throughout lectures or exclusively at the end, and the weight given to retrieval practice exercises in the grading scheme. Regardless of the details, we believe instructors should consider harnessing the power of retrieval practice to enhance memory for the lectures they so painstakingly develop.

Acknowledgment
We thank James M. Edlin for assistance with preparing this article.

Declaration of Conflicting Interests
The authors declared no potential conflicts of interests with respect to the authorship and/or publication of this article.

Financial Disclosure/Funding
The authors received no financial support for the research and/or authorship of this article.

References


Bios
Keith B. Lyle is an Assistant Professor of Psychological and Brain Sciences at the University of Louisville.

Nicole A. Crawford is a PhD candidate in the Department of Psychological and Brain Sciences at the University of Louisville.
Assessment of Problem-Based Learning in the Undergraduate Statistics Course
Christie P. Karpiak
Teaching of Psychology 2011 38:251
DOI: 10.1177/0098628311421322

The online version of this article can be found at:
http://top.sagepub.com/content/38/4/251

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Society for the Teaching of Psychology

Additional services and information for Teaching of Psychology can be found at:
Email Alerts: http://top.sagepub.com/cgi/alerts
Subscriptions: http://top.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Oct 6, 2011
What is This?
Assessment of Problem-Based Learning in the Undergraduate Statistics Course

Christie P. Karpiak

Abstract
Undergraduate psychology majors (N = 51) at a mid-sized private university took a statistics examination on the first day of the research methods course, a course for which a grade of “C” or higher in statistics is a prerequisite. Students who had taken a problem-based learning (PBL) section of the statistics course (n = 15) were compared to those who had taken a traditional format section of the same course (n = 36). PBL students earned higher scores on the examination, taken several weeks after the statistics course had ended, than did students from traditional format courses.

Keywords
problem-based learning, statistics, statistics education

Pedagogies that promote active learning have been recommended by educators for decades (e.g., Bonwell & Eison, 1991; Boyer Commission on Educating Undergraduates in the Research University for the Carnegie Foundation for the Advancement of Teaching, 1998; McKeachie & Kimble, 1950, as cited in McKeachie, 2003). According to cognitive learning theory, active engagement of students in the learning process should correspond with higher motivation in the classroom, deeper processing of course material, and greater retention of information than occurs in traditional lecture (McKeachie & Svinicki, 2006). Indeed, according to some proponents of active learning (e.g., Halpern & Hakel, 2002; Volpe, 1984), traditional lecture is inadequate for most kinds of student learning, due largely to the passive role students take during lecture. Michael (2006) provides a summary of relevant findings from cognitive and learning sciences.

Active learning techniques (e.g., discussions, demonstrations, collaborative activities) have been compared to lecture by researchers from several fields and generally have been found superior (see McKeachie & Svinicki, 2006). However, an array of methods falls under the heading of active learning, empirical support varies across them, and results for Problem-Based Learning (PBL) are more complicated than for other methods (Prince, 2004).

PBL is a form of student-centered instruction that incorporates several active learning techniques and comprises a substantial change from traditional instruction. Students typically work in small groups charged with developing solutions to complex problems. Problems contain sufficient detail to provide direction in gathering information but don’t provide all the information needed to develop a solution. Students must organize the information they have, identify what they need, gather information, consider potential solutions, and communicate the rationale for the final solution they generate (see Duch, Groh, & Allen, 2001). PBL has been used in medical schools for decades, and a prototypical example of PBL is a course developed around consideration of complex medical cases prior to the associated lectures.

Empirical support for PBL is strong for outcomes such as student motivation, teamwork, and self-directed pursuit of learning resources (e.g., Gallagher, 1997; Vernon & Blake, 1993). Much less is known about PBL’s impact on learning of course material, and findings in this area are limited to populations that might differ in important ways from a typical class of undergraduates. The majority of PBL studies suitable for meta-analysis come from medical education (Gijbels, Dochy, Van den Bossche, & Segers, 2005), and most studies outside medicine lack a non-PBL comparison group or do not measure what students have learned.

Strengths of PBL for learning in medical courses include better long-term retention and conceptual understanding of information and stronger understanding of principles that link concepts (Antepohl & Herzig, 1999; Gallagher, 1997; Gijbels et al., 2005; Vernon & Blake, 1993). Results are mixed for the amount of basic knowledge gained from medical courses. Authors of meta-analyses (e.g., Dochy, Segers, Van den Bossche, & Gijbels, 2003) and reviews (e.g., Prince, 2004;
Svinicki, 2007) emphasize the need for ongoing evaluation of PBL’s impact on learning.

The current study addresses learning in the undergraduate psychology statistics course, where motivation is of concern, conceptual understanding and application of concepts are arguably more important than memorization of facts (Keeler & Steinhorst, 1995), and retention of information is desirable for application in subsequent courses.

**Method**

**Implementation**

Statistics in the Behavioral Sciences is a 200-level course with an emphasis on inferential statistics. This course is required for psychology majors and taken by minors and others. A grade of C or better is prerequisite to the majors’ research methods course. A typical class contains 30 students.

After attending a multi-day training workshop, the author changed delivery of her sections of statistics to PBL. The PBL course covers the same topics as the traditional course, consistent with the university’s catalog description, but utilizes cooperative learning groups (3 to 4 students) and learning is initiated by complex realistic problems instead of lectures. Course development was guided by materials from the University of Delaware’s Institute for Transforming Undergraduate Education (ITUE), and the first two problems developed by the author were reviewed by and published on the ITUE’s PBL clearinghouse (https://primus.nss.udel.edu/Pbl/). Table 1 contains practical considerations in implementing a PBL statistics course, including differences between the traditional and PBL courses previously presented at a meeting of the National Institute on the Teaching of Psychology (NITOP) (Karpiak, 2007).

Approximately 6 weeks were spent developing the first set of problems, five or six of which are scheduled into a given semester. The Journal of Statistics Education and related websites also aided in generating ideas. Most problems require students to investigate and reach logic- and data-informed decisions about competing claims. Table 2 contains examples of problem content. Introductory sections of problems contain scenarios to engage students, and most problems use publicly accessible data. An introductory-level text is used and the instructor provides notes to guide reading of the text. Relevant reading is assigned before starting each problem.

**Participants**

On the first day of class in spring 2008, all students (N = 51) in the laboratory for the undergraduate research methods course were administered an examination. Students were psychology majors. Fifteen (29%) had taken the PBL course and the remaining 36 a traditional course. Course format is not listed in course registration materials. A majority of students at this Jesuit University are middle-/upper-middle SES, Catholic, and from northeastern and mid-Atlantic states. Approximately 64% are women, 10% minorities.

**Measure**

Three psychology faculty who teach statistics and/or research methods jointly developed the examination to test students’ knowledge of the statistical information most relevant to research design. We intended items to be difficult, assessing students’ capacity to apply learning from the prerequisite statistics course to the work of the research methods course.

The first section of the examination required students to identify the best statistical test for a given research situation: For example, “collect height and weight data on lots of males and test the hypothesis that weight is related to height.” The answer, in this case Pearson Correlation, was selected from a provided list. These items required recognition of statistical
Table 2. Sample Statistical Topics Addressed With Problems

<table>
<thead>
<tr>
<th>Statistical Topic</th>
<th>Problem Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-normal distributions</td>
<td>Investigation of competing claims about baseball players' salaries</td>
</tr>
<tr>
<td>Normal curve probabilities</td>
<td>Analysis of political claims about tax breaks, &quot;average&quot; income</td>
</tr>
<tr>
<td>Correlation &amp; regression</td>
<td>Contrast between Internet- and valid test-generated IQ scores</td>
</tr>
<tr>
<td><em>t</em> tests</td>
<td>Analyses of competing claims about the quality of public education</td>
</tr>
<tr>
<td>ANOVA</td>
<td>Analysis of the relationship between poverty and other variables</td>
</tr>
<tr>
<td>Nonparametric/chi-square</td>
<td>Investigation of claims of discrimination in mortgage lending*</td>
</tr>
<tr>
<td></td>
<td>Investigation of claims about various dietary supplements</td>
</tr>
<tr>
<td></td>
<td>Investigation of auto safety, effects of various safety equipment*</td>
</tr>
<tr>
<td></td>
<td>Temperature across two centuries by hemisphere</td>
</tr>
<tr>
<td></td>
<td>Voting roll calls in House or Senate by party</td>
</tr>
</tbody>
</table>

Note. Data and the general idea around which the (*) asterisked problems were built came from the DASL database, at http://lib.stat.cmu.edu/DASL/. Student groups are provided with data by the instructor when they reach the relevant parts of the problem. Detailed information about this timing for two problems is found in "teaching notes" on the PBL clearinghouse. The IQ scores problem, developed out of personal interest, uses individual scores made up by the instructor and students are provided with the means and standard deviations from the distributions underlying the Wechsler and the Stanford-Binet IQ tests. Remaining problems were developed from news stories and/or personal interests and use easy-to-find publicly accessible data.

tests and application of knowledge about the variables analyzed via each test. The second section covered statistical inference and null hypothesis testing and required free recall responses to one detailed scenario. The final section contained conceptual and applied multiple-choice questions about statistical power and errors.

Results

An independent samples _t_ test was run on the total score from the examination. PBL students’ scores (_M_ = 11.13, _SD_ = 3.56) were higher than traditional lecture (_M_ = 7.17, _SD_ = 3.26) and the effect size large, _t_ (49) = 3.93, _p_ < .001, _d_ = 1.12.

Additional _t_ tests were run to explore differences in each of the areas of the examination. Differences were significant and effects moderate to large in all sections: selection of the best statistical tests, _t_ (49) = 2.10, _p_ = .041, _d_ = .60; null hypothesis testing, _t_ (49) = 3.77, _p_ < .001, _d_ = 1.08; and power/errors, _t_ (49) = 2.67, _p_ = .003, _d_ = .76.

The university’s student course evaluation form provided a consistent metric for comparison of student experiences of this instructor’s statistics courses from semesters before and after implementation of PBL. Analyses of these forms showed significantly higher overall student ratings of the PBL course, while student ratings of the instructor were stable. A report of these analyses may be obtained from the author.

Discussion

PBL corresponded with better performance than lecture on a statistics examination that was administered approximately 8 weeks after the statistics courses had ended. This is an important finding given the paucity of learning outcome studies of PBL in an undergraduate population, and is interesting in light of inconsistent results in the medical literature regarding PBL’s effect on exam performance. The conceptual nature of this examination was likely better suited to exposing PBL-related group differences than a typical content examination from a statistics class (see Gijbels et al., 2005). Of course, it is not clear whether the stronger performance of the PBL students was due to deeper conceptual understanding of statistics, more flexible application of statistical knowledge, better retention across time, and/or something else.

There are several limitations of this non-experimental study. Most important is the potential role of heterogeneity in the non-PBL condition. All sections of statistics share a common catalog description, but possible contributions to test performance of differences between the faculty members who teach the course, textbooks assigned, time allotted to each topic, and so forth cannot be addressed with this design.

Instructors of the non-PBL courses varied in years of experience, teaching styles, and popularity with students. One received a prestigious teaching award, rendering less compelling any suggestion that the mean differences in test scores might be due to the PBL instructor being a “better teacher.” Furthermore, standard deviations in test scores are not large in the non-PBL group and are comparable between groups. The proposition that differences between instructors might be less important than differences between active and passive pedagogical approaches is not intuitive, but it is consistent with well-replicated findings from the physics literature regarding basic concepts learned in entry-level physics courses (Hake, 1998).

Replication and extension of findings are of primary importance. Refinement of outcome measures or identification of existing validated common measures is also needed. Beyond this, it will be important to identify the specific components of PBL that correspond with differences in undergraduate learning. Statistics educators outside of psychology have studied cooperative groups without complex problems, and others have developed challenging problems but do not use groups. It is not clear which components of the current PBL course contributed to student learning. Likely candidates are the synergistic core elements of collaborative group work and real-world problem solving (see Prince, 2004, for effect sizes associated with these components in the medical school literature), but it is not clear how these components interact with each other, or
whether all are needed for maximal effectiveness. It is also
possible that factors such as the stricter attendance policy or
written work are largely responsible for the measured
differences. A “core elements” approach such as the one
described by Prince (2004), with clear definition of the elements
present in both the PBL and traditional courses, is needed.

Declaration of Conflicting Interests
The author declared no potential conflicts of interest with respect to
the research, authorship, and/or publication of this article.

Funding
Attendance at the Problem-Based Learning workshop was made
possible by a grant from the Dean’s Office, College of Arts and
Sciences, University of Scranton.

Acknowledgments
Thanks to J. Timothy Cannon for his work on the measure and helpful
comments on an earlier version of this article.

References
lecture-based learning in a course of basic pharmacology:
tement in the classroom (ASHE-ERIC Higher Education Report
Boyer Commission on Educating Undergraduates in the Research
University for the Carnegie Foundation for the Advancement of
Teaching. (1998). Reinventing undergraduate education: A blue-
print for America’s research universities. Retrieved from http://
naples.cc.sunysb.edu/Pres/boyer.nlsl/673918d46fbf653e852565ec
0056ff3e/d955b61ffdd590a852565ec005717ae/FILE/boyer.pdf
Effects of problem-based learning: A meta-analysis. Learning and
Instruction, 13, 533-568.
Duch, B. J., Groh, S. E., & Allen, D. E. (Eds.). (2001). The power of
from, what does it do, and where is it going? Journal for the
Effects of problem-based learning: A meta-analysis from the angle
Hake, R. R. (1998). Interactive engagement vs. traditional methods:
A six-thousand-student survey of mechanics test data for intro-
Retrieved from http://dx.doi.org/10.1119/1.18809
Halpern, D. F., & Hakel, M. D. (2002). Learning that lasts a lifetime:
Teaching for retention and transfer. New Directions in Teaching
and Learning, 89, 3-7.
Kariak, C. P. (2007, January). From traditional lecture to problem-
based learning in undergraduate statistics: Why and how. Poster
presented at the National Institute on the Teaching of Psychology,
St. Petersburg Beach, Florida.
Keeler, C. M., & Steinhorst, R. K. (1995). Using small groups to pro-
more active learning in the introductory statistics course: A report
from the field. Journal of Statistics Education, 3(2). Retrieved
from http://www.amstat.org/publications/JSE/v3n2/keeler.html
McKeachie, W. J. (2003). McKeachie’s teaching tips: Strategies,
research, and theory for college and university teachers (11th ed.).
Strategies, research, and theory for college and university teachers
Michael, J. (2006). Where’s the evidence that active learning works?
Advances in Physiology Education, 30, 159-167.
Retrieved from ProQuest Education Journals.
Svinicki, M. D. (2007). Moving beyond “it worked”: The ongoing
evolution of research on problem-based learning in medical educa-
tion. Educational Psychology Review, 19, 49-61.
work? A meta-analysis of evaluative research. Academic Medicine,
68, 550-563.
Zoologist, 24, 433-441.
The Correlator: A Self-Guided Tutorial

Mark L. Mitchell
Janina M. Jolley
Clarion University of Pennsylvania

We describe a free Macintosh® program suitable for use in introductory psychology as well as more advanced classes such as statistics or research methods. The program helps students to distinguish between positive and negative correlation coefficients and to understand the differences between correlation coefficients of different sizes.

Students who do not understand correlation coefficients will not understand the results of psychological research. When these students encounter studies that report negative correlations between variables, they may believe that the studies have found no relation between variables. When they encounter studies reporting a significant positive correlation between two variables, they may believe that the researcher is claiming a near-perfect correspondence between the two variables (Duke, 1978).

If they accept the “claim” of a near-perfect correspondence between the variables, they may form inappropriate overgeneralizations. If, on the other hand, they remember a case inconsistent with this claim, they may reject both the research finding and the validity of psychological research. Thus, a lecture on sex differences could lead some students to embrace gender stereotypes, while leading others to decide that psychological research is invalid (Duke, 1978). Similarly, discussing the validity of an aptitude test may lead some students to decide that psychological research is inaccurate because they know of cases where the test did not predict performance, while leading other students to put undue faith in the test.

Despite the importance of understanding correlation coefficients, there are at least two reasons why instructors may be tempted to skip discussion of correlation coefficients. First, professors cannot easily cover an entire textbook in a single semester. Second, spending considerable class time on the topic often results in only a few students being able to meaningfully interpret correlation coefficients—even when the instructor uses diagrams and examples (Duke, 1978).

To help students understand correlation coefficients without consuming class time, we developed the Correlator—an interactive computer tutorial that helps students explore positive, zero, and negative correlation coefficients. For each type of correlation, students see animations illustrating the relation, develop their own hypotheses, and discover what type of correlation their hypotheses predict. In addition, students match correlation coefficients to simple data generated by a virtual carnival Barker, develop their own data, and answer several questions based on their data.

After students learn about the different types of correlation coefficients, they explore variations in strength of coefficients through the use of dynamic visuals and guided questions. Students compare the relative strength of negative and positive correlations and generate correlations that are stronger than computer-generated ones. Students learn about the median as a way of understanding the strength of various correlation coefficients. The tutorial concludes with an exercise that incorporates Boatright-Horowitz’s (1995) demonstration of Nuttin’s (1985) name letter effect. Once
students complete the tutorial, the computer prints a report documenting when they did the tutorial, how long they spent on it (students typically spend between 20 and 40 min), and the percentage of questions they answered correctly.

Student Acceptance and Learning

In a small pilot study, nine members of one class completed a voluntary, anonymous survey. All who responded either agreed or strongly agreed that working through the tutorial was more interesting than trying to learn the information from reading a textbook. Likewise, all students either agreed or strongly agreed that the tutorial had improved their understanding of correlation coefficients.

To better assess whether the Correlator aided student learning, we performed two additional studies. Both samples consisted of students enrolled in introductory psychology classes. The first study found that students who completed the Correlator for extra credit (n = 28) scored an average of 10 points higher on the subsequent exam over research methods and social psychology than those who did not (n = 39). The second study compared a class introduced to the Correlator (n = 36) to a class that was not (n = 38). Although the no-Correlator class averaged 6 percentage points higher on the final exam, the Correlator class averaged 10 percentage points higher on the four final exam questions that involved interpreting correlation coefficients.

Conclusions

As would be expected from decades of data documenting that computer-assisted instruction is effective (Desberg, 1994), students completing the Correlator tutorial better understand correlation coefficients. However, contrary to research showing students sometimes find computer-assisted instruction boring (Desberg, 1994), students seem to enjoy going through the tutorial.

References


Notes

1. A preliminary version of this article was presented at the annual meeting of the American Psychological Association, Toronto, August 1996.
2. We thank David Brooks and Helen Brooks for their assistance in developing the Correlator. We also thank Janet Morahan-Martin, David J. Pittenger, Jeanne M. Slattery, and Randolph A. Smith for comments on an earlier version of this article.
3. Anyone wishing to receive a copy of the Correlator tutorial can download it from http://psy1.clarion.edu/mm/download/correlator.
4. Send correspondence to Mark L. Mitchell, Department of Psychology, Clarion University, Clarion, PA 16214; e-mail: mitchell@mail.clarion.edu.
Evaluation of an Interactive Tutorial for Teaching the Central Limit Theorem

Christopher L. Aberson
Dale E. Berger
Michael R. Healy
Diana J. Kyle
Victoria L. Romero
Claremont Graduate University

In this article, we present an evaluation of a Web-based, interactive tutorial used to present the sampling distribution of the mean. The tutorial allows students to draw samples and explore the shapes of sampling distributions for several sample sizes. To evaluate the effectiveness of the tutorial, 111 students enrolled in statistics or research methods courses used either the interactive tutorial or attended a lecture and a demonstration on the sampling distribution of the mean. Students in both groups improved from pretest to posttest and no statistically significant differences between improvement scores were found between groups. Additionally, students rated the tutorial as easy to use and understand. In this study, we provide evidence that an Internet tutorial can be comparable in effectiveness to standard lecture or demonstration techniques.

The central limit theorem states that the sampling distribution of the mean for any population, given an adequate sample size, will approximate a standard normal distribution. Understanding sampling distributions is essential to comprehending many core statistical techniques. Unfortunately, students often fail to grasp sampling distribution concepts as presented in statistics textbooks (Dyck & Gee, 1998; Zerbolio, 1989). Consequently, students become confused with advanced topics (Howell, 1997). The sampling distribution of the mean allows one to calculate the probability of obtaining a sample mean given the sample size, population variance, and population mean. Researchers use this probability to assess the validity of null hypotheses and construct confidence intervals. Unfortunately, the teaching of statistical procedures often emphasizes working through problems in a “cookbook” fashion, focusing on mechanics instead of the logic of applications (Garfield, 1995). As a result, students may learn how to reject hypotheses based on a comparison of calculated values to tabled values with little understanding of the reasoning behind the test.

The Web Interface for Statistics Education (WISE) project provides instruction that addresses these shortcomings. The WISE project has created several Web-based tutorials that require only a Java-enabled browser. In this article, we present an evaluation of an interactive, Internet-based tutorial to assist students in learning about sampling distributions. The tutorial, found at http://wise.cgu.edu under “tutorials,” consists of a paper-based assignment that guides students through the use of an interactive applet. The applet allows students to compare the population distribution with sampling distributions for samples of various sizes drawn from several different population distributions, although this assignment uses only normally distributed populations. The sampling distributions applet simulates the results of drawing many random samples. This tutorial assumes an audience that has basic knowledge of sampling, means, and normal distributions. For those students who have completed instruction on these topics, the tutorial can serve to replace some in-class instruction on sampling distributions.

The assignment begins with a problem statement. Students read that they will investigate life satisfaction as measured by a scale with a reported mean of 0.50 and standard deviation of 0.20. Using the applet, the student simulates drawing a sample of 100 scores. The applet shows the distribution of the sample cases and the sample mean. The student records the sample mean and notes whether it falls within 0.05 points of the population mean. The student draws 9 more samples of 100 scores and records each sample mean. Then, the student examines the actual sampling distribution of means for samples of 100 scores and estimates the propor-
tion of sample means falling within 0.05 of the population mean. Next, the student uses standard \( z \)-score formulas to calculate, by hand, the proportion of sample means expected to fall within the same range around the population mean. The student then relates the observed distribution of sample means to the calculations and estimates by comparing the calculated percentage of sample means expected to fall within a certain range to the observed proportion of sample means actually falling within that range.

The assignment continues this process for samples of 25 scores and then for samples of 5 scores. After drawing samples of 5 scores and comparing results, the student evaluates the following statement: “For any population, the best estimate of the mean is the sample mean—therefore, it shouldn’t matter what size sample I use. I’ll use a sample of \( n = 5 \) as this will save a good deal of time and money.” The tutorial instructs students to respond to the statement, using information from the exercises that they have just completed.

Method

Participants

One hundred and eleven students, 34 men and 77 women, enrolled in introductory statistics (three sections), intermediate statistics (one section), and research methodology (one section) courses participated in the study as part of regular instruction on sampling distributions. Participants in the study were students from two colleges, a large state university (\( n = 73 \)) and a large community college (\( n = 38 \)). The study included 12 freshmen, 28 sophomores, 29 juniors, 33 seniors, and 9 graduate students. The majority of the students participating in the evaluations were of traditional college age. The course instructor randomly assigned students to the lecture and demonstration or the tutorial within constraints imposed by scheduling restrictions (e.g., limits in the number of available computers).

Procedures

Students attended either a lecture and demonstration on sampling distributions or used the tutorial assignment. The tutorial group (\( n = 55 \)) received a packet with a \( z \) table, a review of calculations of \( z \), an overview of the applet, and an exercise guiding use of the applet. These students went to a computer lab and worked independently on the tutorial. Students received assistance navigating the tutorial but did not receive assistance with statistical concepts.

The lecture group (\( n = 56 \)) attended a lecture and demonstration of the sampling distribution of the mean. The demonstration presented a population of 20 to 35 exam scores from which students drew samples of varying sizes. Exam scores were written on slips of paper and drawn from a paper bag. For each set of samples, the class examined how closely the sample means approximated the population mean and the proportion of sample means falling within a certain distance of the population mean compared to the proportion that would be expected to fall within that range. The lead author wrote the lecture and demonstration used for all lecture conditions.

Evaluation Materials

Comprehension. Quizzes measured knowledge of sampling distributions before and after instruction. The pretest and posttest both contained questions assessing knowledge of calculation procedures (e.g., probability of obtaining a specific mean value), theory (e.g., relation between sample size and standard error), definition (e.g., identify sample and population means), and application (e.g., decisions involving sample sizes for a research problem). Students completed the pretest measures at the end of the class meeting immediately prior to the lecture or tutorial session. Students completed the posttest measure immediately following the lecture or tutorial.

Student ratings. Students responded to questions regarding the amount that they believed they learned about the topic; how useful they viewed the lecture or tutorial; how clear they found the explanation of statistical concepts; how easy the lecture or tutorial was to understand; the quality of the explanation of statistical concepts; comfort with computers; and demographic information such as sex, class level, and grade point average.

Results

Equivalence of Groups

Those students completing both the quiz and student ratings sections proved to be similar with regard to several variables. No statistical difference between groups existed for grade point average, \( t(99) < 1 \); sex, \( \chi^2(1, N = 111) = 1.69, p = .12 \); class level, \( \gamma = -.20, p = .15 \); or comfort with computers, \( t(108) = 1.56, p = .15, \eta^2 = .02 \).

Quiz Results

Scores on pretest and posttest quizzes indicated that students using the tutorial and students attending the lecture and demonstration learned comparable amounts. Overall, average performance improved by 2.3 points on the 9-point quiz, \( F(1, 109) = 148.5, p < .001, \eta^2 = .58 \). Scores on pretest quizzes showed that the tutorial group (\( M = 4.66 \)) performed slightly, although not significantly, better than the lecture group (\( M = 4.21 \)), \( F(1, 109) = 1.30, p = .26, \eta^2 = .01 \). Scores on the posttest were similar for tutorial (\( M = 6.73 \)) and lecture (\( M = 6.79 \)), \( F(1, 109) < 1 \). No interaction between pretest and posttest quiz score and learning condition was present, indicating that there were no significant differences in the amount each group improved (Tutorial \( M = 2.07 \), Lecture \( M = 2.58 \)), \( F(1, 109) = 1.72, p = .19, \eta^2 = .02 \). Although the tutorial group showed slightly (but not significantly) less improvement, the appearance of this difference may result from the fact that the tutorial group’s pretest was slightly higher than the pretest of the lecture group.
dence intervals around improvement scores for each group indicated that the tutorial group (95% CI = 1.48 to 2.67) and the lecture group (95% CI = 2.09 to 3.06) both improved significantly.

Student Ratings

Students’ ratings provided additional evidence for the similarity of the two learning conditions. Forty-three additional students provided ratings of the tutorial or lecture, raising the overall sample size to 154 (tutorial, n = 83; lecture, n = 71). Students’ ratings did not differ between learning condition for ratings of the usefulness of the presentation of topic, \( \gamma = -0.04, p = .80 \); the amount they had learned, \( \gamma = 0.01, p = .96 \); ease of understanding, \( \gamma = 0.13, p = .32 \); or clarity of explanation of statistical concepts, \( \gamma = -0.03, p = .85 \). The majority of students in both learning conditions rated the presentation as being useful, believed that they had learned “a lot,” found the instruction easy to understand, and rated statistical explanations as “clear.”

Discussion

Pretests and posttests of knowledge demonstrated that the Web-based tutorial was comparable in effectiveness to a good lecture or demonstration in fostering learning about a core statistical concept. An implication is that this tutorial could replace some in-class presentations, freeing class time for discussion or other activities that cannot be incorporated easily into a computerized module. The tutorial provides some important advantages over the lecture format. Perhaps most important, the interactive tutorial gives students substantial control over the learning process. The student can access the tutorial at any time, proceed at any desired pace, stop at any time, redo portions of the module, and so on. It is also important to note that most students rated the tutorial as easy to use and understand, and they perceived the tutorial to be effective in teaching the topic. As students gain experience with computer-based tutorials, we may see even greater acceptance of this medium.

The sampling distribution tutorial incorporated several techniques for enhancing student learning. These techniques include providing authentic challenges to student understanding, presenting multiple perspectives, encouraging reflective thinking, and addressing misconceptions. Authentic challenges (e.g., Park & Hannaﬁn, 1993) are provided through materials that go beyond calculation of statistical values. Multiple perspectives (e.g., Mayer & Anderson, 1992) are presented through problems that require interactive manipulation, sampling, estimation, and interpretation in addition to calculation. Reflective thinking (e.g., Hoffman & Ritchie, 1997) is encouraged through questions that force students to stop at various points in the tutorial and respond to questions about results. Student responses to questions that represent common misinterpretations of concepts allow for active confrontation of misconceptions (e.g., Garﬁeld, 1995). As demonstrated by the sampling distribution tutorial, many important principles of good instruction can be incorporated into Web-based tutorials. Given the encouraging results of this assessment, it is clear that interactive computer-based tutorials can provide an effective supplement, or even replacement, for traditional classroom lectures. However, it is important to test the effectiveness of educational software, as neither users nor instructors can be expected to make valid subjective judgments regarding effectiveness (cf. Jolicoeur & Berger, 1986).

References


Notes

1. The Web Interface for Statistics Education project is supported by grants from the Mellon Foundation and Claremont Graduate University.
2. Send correspondence to Christopher L. Aberson, who is now at the Department of Psychology, Humboldt State University, Arcata, CA 95521; e-mail: cla18@axe.humboldt.edu.
"Lectlets" Deliver Content at a Distance: Introductory Statistics as a Case Study
Russell T. Hurlburt
Teaching of Psychology 2001 28: 15
DOI: 10.1207/S15328023TOP2801_04

The online version of this article can be found at:
http://top.sagepub.com/content/28/1/15

Published by:
http://www.sagepublications.com

On behalf of:
Society for the Teaching of Psychology

Additional services and information for Teaching of Psychology can be found at:

Email Alerts: http://top.sagepub.com/cgi/alerts

Subscriptions: http://top.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Jan 1, 2001

What is This?
A lectlet is a short Web streamed-audio lecture synchronized to an interactive text–graphics display. Lectlets can be used either as distance-education courseware or as enhancements to traditional classroom education. In this study, I compared lectlet-based distance education to traditional education for 4 semesters by simultaneously teaching 2 sections of introductory statistics: one lectlet based and the other in traditional format. Students in each format visited the other format and compared them. Student evaluation and course outcomes showed that (a) lectlets effectively presented content, (b) lectlet-based format was better than the traditional course in 6 specific ways, (c) the traditional format was rated as a better overall learning environment, and (d) lectlets were valuable as ancillaries to the traditional format. Lectlets may be useful in any content-oriented course.

Distance education is the “fastest growing form of domestic and international education” (McIsaac & Gunawardena, 1996, p. 403) as students and institutions are lured by the convenience of courses with fewer constraints of location or time (Guernsey, 1998). Furthermore, many traditional classroom courses (25% according to Green, 1997) include Internet material similar to distance-education courseware. However, the understanding
of distance-education and at-a-distance traditional course enhancement methods has not kept pace with their rapid increase. For example, the Report of the Task Force on Distance Education (1992) noted that it is a mistake simply to take live-classroom material and transfer it to another medium such as video: Distance education “is a form that demands that the design, production, and distribution of programs be specifically geared to distance-education learning objectives and take into account all of the special needs and requirements of individual programs” (p. 5). However, Boling and Robinson (1999) noted that there has been little theoretical or empirical evaluation of the media and methods developed to present distance-education material. McIsaac and Gunawardena concluded that “it is time … to examine … how best to incorporate media attributes into the design of effective instruction for learning” (p. 421).

Although McIsaac and Gunawardena (1996) targeted distance education in general, their conclusion applied in particular to the teaching of introductory statistics. To date the only published account of a distance-education statistics course is McCollum’s (1997) brief description of a course that combined a Web threaded discussion with an electronic chat room. McCollum reported that the distant students performed better than the traditional classroom students but provided no data.

There are a few accounts of at-a-distance portions of traditional statistics courses. Varnhagen, Drake, and Finley (1997) found that “the communication components of the Internet laboratory [were] more useful than the information components” (p. 275) for an intermediate statistics course. Portier, Hermans, Valcke, and van den Bosch (1997) used a computer-based “electronic workbook” as an ancillary to traditional instruction. Portier and van Buuren (1995) described an “interactive learning course development environment,” a flexible, computer-based method of delivering a variety of course materials. They delivered the working-group portion of a statistics course using this method. Martens, Valcke, Portier, Weges, and Poelmans (1997) described a home computer descriptive statistics segment of a statistics course. Couch (1997) created an Internet site that posted lecture notes and PowerPoint® slides; students reported that they valued having the notes before a lecture to get an idea of what the lecture would cover and to organize their comprehension.

In summary, instructors have little evidence about how to create entire distance-education courses or how to create effective distance-learning enhancement for students in traditional courses. This article describes and evaluates one method of at-a-distance content delivery that is useful both for entire distance-education courseware and for traditional course enhancement. This method may be useful for a variety of courses; I evaluate it here for introductory statistics.

**Lectlets**

I decided in 1996 to create a distance-education statistics course to answer this question: Is it possible to construct a distance-education course equal to or better than our traditional classroom course? At the outset, I had no preconceptions as to the medium or method of distance-education presentation other than that the course had to be available at remote sites.

After considering many alternatives, I created a new medium called the lectlet, a short Web audio lecture (my voice delivered as streamed audio) synchronized to an interactive text–graphics display on the computer monitor. Note that the term lectlet is an analogy to the computer programmer’s use of the term applet, a small application program. Because lectlets seemed to be a potentially effective way to teach content-laden material such as statistics, I constructed a series of lectlets for the introductory statistics course.

There were 2 to 4 lectlets for each chapter of the class textbook (Hurlburt, 1998). Each of the 39 lectlets was six to eight “pages” long. Each page involved a 30-sec to 9-min streamed RealAudio lecture and an interactive synchronized visual display on a portion of the computer screen that I called the blackboard. The blackboard provided both text and figures. Blackboard texts were short, as if on a real blackboard, and included histograms, scatterplots, and so on. Blackboard texts were frequently color coded to specific portions of the figures.

The first three or four pages of each lectlet presented an interactive review of the previous lectlet. Students responded to a question or series of questions by using the keyboard to enter their answers in text-entry cells on the blackboard. The next page provided immediate feedback.

Seven of the lectlets provided interactive or animated demonstrations. For example, the lectlet dealing with random samples displayed a normal distribution with four Monte Carlo samples. Students could click “take another sample” repeatedly to gain a sense of the characteristics of such samples.

The 39 lectlets were entirely asynchronous: They were accessible at any time, in any order, and were repeatable. A custom interface gave students control over the lectlets. By clicking appropriate buttons on the interface, students could perform a variety of operations on a lectlet: They could advance to the next page or skip forward or backward to any other page; enable or disable the automatic launching of the audio; pause, stop, rewind, or replay the audio and adjust its volume; display, hide, or print a word-for-word transcript of the audio; print the lectlet blackboards; advance to the next lectlet; display the course syllabus or schedule; and obtain help.

**Evaluation of the Lectlet Format**

To evaluate the lectlets as a method of distance education, I taught simultaneous pairs of sections of an introductory statistics course. One section of each pair used the traditional (live in the classroom) format, whereas the other section used the lectlets along with a listserv that allowed students to interact with each other and with me. The two sections of each pair were otherwise identical, using the same textbook material (Hurlburt, 1998, chapters 1–13; 16–18), statistical software (Hurlburt, 1995, 1997), calendar, exams, and so on. Traditional students took the exams, which were a combination of multiple-choice and computational items, in the classroom. Lectlet-based students took the exams at a proctored site (either on campus or at a distant location) during a 4-hr window that overlapped the traditional students’ exam period. I have taught this pair of sections three times, and the lectlets remained nearly identical except for minor correc-
tions such as typographical errors. This course was a requirement for psychology majors, who comprised roughly two thirds of the students, and for majors from other departments such as nursing. The initial enrollments were 51, 40, and 40 students in the traditional sections and 8, 13, and 12 students in the distance-education sections.

In addition, a colleague also taught a pair of traditional and distance-education statistics sections. His traditional section used the same textbook and covered the same material as did the lectlet section, but he used his own classroom method and personal teaching style. His distance-education section used the lectlets previously described. He used his own exams, which were identical in both sections. The initial enrollments were 42 students in his traditional section and 19 students in his lectlet-based section.

Results

Six procedures evaluated the lectlets: lectlet-by-lectlet feedback, anonymous course evaluation, direct comparison, course completion rates, course grades, and comments from the colleague.

Lectlet-by-Lectlet Feedback

The last page of each lectlet included an electronic form to submit comments on that lectlet. The infrequent comments fell into three categories: positive comments, reports of mistakes (usually typographical errors), and reports of problems with lectlet delivery (e.g., the network was slow).

Anonymous Course Evaluation

Lectlet-based students completed an anonymous electronic evaluation at midterm in three semesters and near the end of the course in the fourth. Twenty-eight of the 43 students still enrolled at that time provided demographic and computer-equipment information, narrative commentary, and Likert-scale ratings.

The 11 narrative commentary items revealed that nearly all students liked the lectlets and the fact that they were available 24 hr per day. The RealAudio stream was successful: Students with 28.8K modems and Pentium computers reported that the audio presentation generally began within 6 to 10 sec of their requesting it. Students frequently commented that the lectlets enhanced the textbook: Hearing the pronunciations and voice inflections aided their comprehension of new material. The division of the lectlets into pages made it easier to find particular parts of a lectlet for review. They liked the ability to pause, rewind, and replay the audio, and used them frequently. Otherwise, they took the lectlet interface for granted, which is evidence that it worked well.

Students commented positively about the interactive review pages of the lectlets. Their incorrect answers to review items convinced them to reconsider aspects of the material they thought they understood.

Table 1 presents the responses to the lectlet-based course evaluation Likert-scale items. All students reported that they always or usually could access the course materials on demand. Nearly all students reported that they always or usually completed the review items.

Students varied in their use of the word-for-word transcripts of the audio, a feature created primarily for special-needs students (e.g., those with learning disabilities or for whom English was a second language). A native speaker of Chinese reported that the transcript of the audio was useful for mastering statistics and also for improving her mastery of English. Many students without special needs also found the transcripts to be extremely useful; nearly all students used them at least occasionally. Four students reported that they printed every transcript for use during the lectlet and reference later. The transcript also facilitated the search process when a student wished to replay a particular portion of a lectlet and provided a useful backup in case of audio difficulty.

Table 1. Course Evaluation Ratings by Lectlet-Based Students from Four Semesters

<table>
<thead>
<tr>
<th>Evaluation Statements</th>
<th>Rating Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>I could access the materials (lectlets, labs, and quizzes) when I wanted to</td>
<td>Always</td>
</tr>
<tr>
<td>I read the assigned chapters</td>
<td>13</td>
</tr>
<tr>
<td>I access the assigned lectlets</td>
<td>22</td>
</tr>
<tr>
<td>I make an honest effort to fill in the blanks in the review portions of the lectlets</td>
<td>19</td>
</tr>
<tr>
<td>When I access a lectlet, I use the audio portion</td>
<td>15</td>
</tr>
<tr>
<td>When I access a lectlet, I use the lectlet transcripts</td>
<td>20</td>
</tr>
<tr>
<td>When I access a lectlet, I use it completely to the end</td>
<td>7</td>
</tr>
<tr>
<td>I access the same lectlet more than once for review/relearning purposes</td>
<td>19</td>
</tr>
<tr>
<td>I generally access the lectlet before I read the chapter</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>9</td>
</tr>
</tbody>
</table>

Note. N = 28. Data are from three semesters taught by the author and one semester taught by a colleague.
In general, the students reported that the lectlets did clearly present the material at a distance. Nearly all students reported using the lectlets to completion although they knew that the lectlets would not provide any information not in the textbook. Most students reported that they at least occasionally used the same lectlets more than once.

**Direct Comparison**

To compare the lectlets and lectures directly (same instructor and same material), I invited all students in my traditional sections to access one or more of the lectlets, and 67 of 94 (71%) did so. I also invited all students in the lectlet-based sections to visit one or more traditional classes, and 8 of 19 (42%) did so. The smaller percentage of participating lectlet-based students reflects the fact that most lectlet-based students had employment that conflicted with the class time. My colleague’s sections could not participate in this direct head-to-head comparison because the same-instructor principle was violated: He created and delivered the lectures in his traditional section, whereas I had created and delivered the lectlets in his lectlet-based section. All 75 students who made this direct comparison received extra credit for providing narrative responses to seven open-ended questions and six items on a 7-point Likert-scale ranging from 1 (strongly in the class) to 7 (strongly on the Web). These six items appear in Table 2.

Most of the 75 students preferred all aspects of the traditional classroom: They thought that it was clearer, that they were more involved with the material, and they preferred it overall. However, most thought the lectlet-based learning was more convenient. Students were neutral about the pace of the formats.

Table 2 lists the frequencies and means for the 8 lectlet-based students who visited the traditional classroom. These data and comments from the other evaluation procedures suggest that the lectlet-based students, like their traditional counterparts, found the traditional classroom more involving and a better learning environment. However, distance students generally preferred the lectlet format despite the fact that they judged the traditional classroom to be a better educational experience.

Table 2 shows an exception to the previous generalization. Although the majority (33 + 22 = 55 of the 75 respondents) were of the strong opinion that they learned better in the traditional classroom, a minority (2 + 3 = 5, including 4 traditional students) strongly held the opposing opinion. Any interpretation of these direct comparisons should recognize that students did not have equal experience with the two formats. Traditional students typically visited only one or a few lectlets, and lectlet-based students typically visited only one or two classes. If students had more experience with both formats, their comparisons might differ.

**Course Completion Rates**

Considering the courses taught both by the author and the colleague, there was no practical or statistically significant difference between the course completion rates in the two formats, \( \chi^2(1, N = 225) = .09, p > .05 \). Of the 173 students who enrolled in one of the four traditional sections, 116 (67%) completed the course. Similarly, of the 52 students who enrolled in one of the four distance-education sections, 36 (69%) completed it.

**Course Grades**

All sections of the course included five 75-min exams and a cumulative final exam, all graded blindly. Although the exams were identical in both formats within each semester, the exams were not identical across semesters. Therefore, I converted the point totals (on which the course final grades were based) to \( z \) scores based on the combination of both formats within each semester, and then aggregated the \( z \) scores within formats across the four semesters. There was no significant difference in the means of the \( z \)-transformed point totals between the traditional format (\( M = -.03, SD = 1.00, N = 116 \)) and the lectlet-based format (\( M = .09, SD = .96, N = 36, t(150) = .62, p > .05 \)). On an A = 90, B = 80, and so forth grading scale, the average lectlet-based student’s point total was thus approximately 1 point higher than the average traditional student’s, a difference that is neither statistically nor practically significant.

### Table 2. Direct Comparison Ratings by All Students from Three Semesters

<table>
<thead>
<tr>
<th>Evaluation Statements</th>
<th>All Students</th>
<th>Lectlet-Based Students&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3 4 5 6 7</td>
<td>1 2 3 4 5 6 7</td>
</tr>
<tr>
<td>I think the material is covered more clearly</td>
<td>19 23 12 15 1</td>
<td>4 1 2 1 2</td>
</tr>
<tr>
<td>I think I learn the material better</td>
<td>33 22 6 9 0 2</td>
<td>1 5 1 1 1</td>
</tr>
<tr>
<td>I think the pace is better</td>
<td>7 10 8 29 9 3</td>
<td>1 1 3 1 2</td>
</tr>
<tr>
<td>I think the convenience is better</td>
<td>4 3 2 8 7 15 36</td>
<td>1 2 5 6 3.8 1.06</td>
</tr>
<tr>
<td>I feel more involved with the material</td>
<td>25 13 8 10 4 3</td>
<td>1 2 2 2 2</td>
</tr>
<tr>
<td>Overall, I prefer</td>
<td>22 24 12 10 0 4</td>
<td>1 2 1 1 3</td>
</tr>
</tbody>
</table>

Note. \( N = 75 \). Ratings were based on a scale ranging from 1 (strongly in the class) to 7 (strongly on the Web). All courses were taught by the author. <sup>a</sup> \( n = 8 \).
This was a field study with students choosing to participate in the traditional or distance-education format, so it is possible that the two groups differed initially on many different variables (e.g., motivation, independence, level of simultaneous employment responsibilities, number of children or other caretaking responsibility, and computer literacy). No attempt was made in this study to control or measure those important variables. Furthermore, students in the two formats may have differed on initial level of mathematical skill. The first exam can serve as a rough indication of incoming mathematical skill because many items tested basic algebra skills, summation notation, and probability concepts. This is, however, only a crude measure of mathematical aptitude because it is confounded with the beginning concepts in statistics. Using transformations as before, there was no statistically significant difference, $t(150) = 1.10, p > .05$, in the first exam scores in the aggregate of the four semesters between the traditional students ($M = -.05, SD = 1.02$) and the distance-education students ($M = .16, SD = .87$). Again, on a $A = 90$ points grading scale, the average lectlet-based student's first exam score was approximately 2 points higher than the average traditional student's. Thus, there were not statistically significant or practical differences between the two formats on either the total points earned in the course or the first exam score.

**Comments From the Colleague**

The colleague is an excellent teacher, having received the highest college-wide and university-wide excellence-in-teaching awards. He is quite student oriented, having been advisor to Psi Chi and the University Psychology Club for many years. By the second day of class he knew every student in all of his classes (75 or so students in total) by name. Furthermore, he is very familiar with distance education, having taught introductory psychology courses by videoconferencing (simultaneous live classroom and remote sites) for four semesters.

Cognitively, he thinks that his lectlet-based students learned the statistical content at least as well as did his traditional students, but he acknowledges that this outcome may have been due to initial group differences (his distance students outperformed his traditional students on the first exam). He believes that the lectlet-based course fostered independent learning skills better than traditional courses do, and that this independent-learning advantage might outweigh his perception that the lectlets may not be quite as effective in presenting content as a good traditional lecture. Thus, he thinks that the lectlet-based, distance-education format was for some students pedagogically superior to traditional education, reversing the opinion of distance education, having taught introductory psychology courses by videoconferencing (simultaneous live classroom and remote sites) for four semesters.

Affectively, he found it strange to have so little personal knowledge of the students in the lectlet-based section, in contrast to his traditional teaching in which he knew each student's strengths and weaknesses and observed each student's acquisition of the new skills. That the lectlet-based students learned the material without his "getting to see it happen" seemed almost magical. He was surprised by the strength of the students' positive feelings about the course (typically expressed in casual e-mails), particularly because he felt little involvement with the lectlet-based students.

**Discussion**

This study evaluated the lectlet, a new format for teaching distance-education courses. Nearly all students had distinctly positive reactions to the lectlets.

The lectlets were apparently better than traditional lectures in six ways. First, the lectlets provided a better review of the immediately preceding material than is possible in the classroom. The interactive review questions on the first several pages of each lectlet allowed each student the opportunity to answer every review item. In contrast, traditional lectures frequently begin with a series of review questions, but only one student can answer each question, whereas the other students may or may not be involved.

Second, the lectlets gave students more control over the material than is possible in the traditional classroom. The ability to pause, rewind, and replay portions of the lectlets is far superior to the opportunity to ask a traditional lecturer to repeat something. Third, students can revisit entire lectlets, which is not usually possible in the classroom. Although traditional lectures can be audiotaped, there are three reasons that revisiting a lectlet may be superior to replaying an audiotape: (a) the lectlet is expressly designed to be replayed; (b) the replayed lectlet is exactly the same as the original, auditorily, visually, and interactively; and (c) the lectlet's pagination and transcript make it easier for the student to scan through a lectlet to find a particular passage to replay.

Fourth, lectlets are asynchronous and potentially always available, whereas classroom lectures occur only at specific times. Fifth, the lectlet transcripts were surprisingly useful for some students. It is impossible to create a useful transcript in the traditional classroom. Some traditional instructors distribute lecture notes, but that is not the same as a word-for-word transcript of an entire lecture for which the audio is readily available. Sixth, the use of lectlets may foster the acquisition of independent learning skills.

Despite these advantages of the lectlets, the traditional course was, in the opinion of both traditional and lectlet-based students, a better educational experience. One might argue that I have been teaching traditional courses for 25 years, whereas this is my first lectlet course. However, I have been creating computer-based demonstrations of statistical concepts for 20 years. There are, to be sure, a few ways that the lectlet-based course might be incrementally improved but none seems substantial enough to alter the overall conclusion that for most students the traditional course is a better educational experience.

A minority of students, however, believed strongly that the lectlet experience was better than the traditional format, either because they found the review pages of the lectlets to be particularly effective or because they had difficulty attending class regularly. Furthermore, there are those students for whom participation in the traditional classroom is simply not possible.

Thus, the primary conclusion is that the lectlets effectively deliver statistical content at a distance. The lectlet
format may also be useful for other course contents, but such a generalization is limited by the fact that this statistics course was designed to convey a substantial amount of information and exercises a number of skills. Other kinds of courses emphasize different goals, such as creative writing skills, persuasion, group process, giving and receiving criticism, and so on. The lectlet method may or may not be effective there.

I make three additional observations. First, the creation of a series of lectlets is a large endeavor. It required approximately 2,000 hr to create the lectlet interface and the 39 lectlets (perhaps 35 hr per lectlet once the interface was completed). A development time of 2,000 hr, although substantial, is not unreasonably large by comparison to other distance-education formats. For example, Eamon (1999) concluded that “the cost of developing multimedia materials can be staggering” and that administrators “vastly underestimate” (p. 205) the costs of technology and distance education. He cited time estimates ranging from 18 to 300 hr and direct cost estimates of up to $50,000 to create 1 hr of multimedia instructional material.

Second, it might be more accurate to refer to the lectlet-based course as a “convenience education” course rather than a distance education course. Nearly all of the lectlet-based students lived in the same city as the university (although a few were thousands of miles away), and many were simultaneously taking other traditional courses in the university. This is not an uncommon phenomenon; Guernsey (1998), for example, reported that more than 80% of students enrolled in the University of Colorado’s distance-education courses were also taking traditional classroom courses. Students selected the lectlet-based format for a variety of reasons (work responsibilities, caretaking responsibilities, or schedule conflicts), most of which had more to do with convenience than with distance.

Third, the lectlets were effective tools as adjuncts to the traditional classroom. Narrative comments on the direct comparisons as well as informal comments indicate that some traditional students used the lectlets frequently for review and to make up for classes missed due to illness or other reason. The enthusiasm that these students had for the lectlets was striking, especially considering that these opinions were not directly solicited. Thus, those students answered the question, “which is better: traditional lectures or lectlets?” with a clear “both together.”

References


Notes

1. I thank Chris Heavey, my colleague who taught the pair of additional sections, for his participation and for comments on earlier drafts of this article.

2. For links to lectlets described here and to other information about the course see http://www.nevada.edu/~russ and follow the “distance education statistics” link.

3. Send correspondence to Russell T. Hurlburt, Department of Psychology, University of Nevada, Las Vegas, NV 89154–5030; e-mail: russ@nevada.edu.
Computer and Software Use in Teaching the Beginning Statistics Course
Albert E. Bartz and Marisa A. Sabolik
*Teaching of Psychology* 2001 28: 147
DOI: 10.1207/S15328023TOP2802_16

The online version of this article can be found at:
http://top.sagepub.com/content/28/2/147

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Society for the Teaching of Psychology

Additional services and information for *Teaching of Psychology* can be found at:

- **Email Alerts**: http://top.sagepub.com/cgi/alerts
- **Subscriptions**: http://top.sagepub.com/subscriptions
- **Reprints**: http://www.sagepub.com/journalsReprints.nav
- **Permissions**: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Apr 1, 2001

What is This?
Computer and Software Use in Teaching the Beginning Statistics Course

Albert E. Bartz
Marisa A. Sabolik
Concordia College

We report the results of a survey examining computer-assisted data analysis in undergraduate psychology departments. Although 90% of the departments offered computer-assisted data analysis somewhere in their curriculum, one third did not use this approach in their beginning-level statistics course. We discuss the implications of the findings and suggestions for future research.

Computer use in psychology departments has steadily grown with the increased availability of microcomputers and appropriate software. One of the earliest surveys on computer use in psychology showed that 50% of respondents reported using computers in statistics and experimental methods courses (Castellan, 1982). Several years later, Couch and Stoloff (1989) documented increasing usage showing microcomputer use for research methods and statistics at 71% and 66%, respectively, of a national sample of psychology departments. More recently, Alder and Vollick (1999) reported that 79% of North American undergraduate departments indicated that "computers were used to some extent in their departments' mandatory statistics course(s)" (p. 2).

Those who teach the beginning statistics course would be especially interested in such survey results, given that there is
not complete agreement that computer software should be used in the beginning statistics course (Dillon, 1999). However, it has been difficult to determine usage in past surveys because the results have not been reported for the beginning statistics course separately. In addition, one or more survey respondents might indicate using a computer for classroom demonstrations, simulations, or data displays, without the students being involved in using computers for data analysis. To avoid these points of confusion, we included both computer and software use in our survey rather than computer use alone. We designed this study to determine the extent of computer usage in the beginning statistics course. A second purpose was to tabulate the variety of statistical software used.

Method

Sample

We sent a cover letter and a short questionnaire on an oversized postcard to psychology department chairs at a sequential sample of colleges and universities. The sample consisted of every fifth school from the CollegeSource Online® Web site that listed an undergraduate major in psychology (N = 1,123). We omitted 21 schools from the sample of 224 because of an improper address format and sent the questionnaire and letter to the remaining 203 schools.

Instrument

Questions on the 9.5 × 21.5 cm self-addressed, stamped postcard asked respondents if they offered an introductory statistics course in their department and, if so, whether they used computer-assisted data analysis in the course. Additional questions requested information regarding text materials, brand of software, and instructional and practice mode (regularly scheduled laboratory periods, class time, students’ own time). The final item asked those who did not use computers in the beginning course to identify if and where in the curriculum they did use computers for data analysis.

Results and Discussion

Of the 203 surveys mailed, 88 (43%) were returned. We then contacted by phone a random sample of 30 of those that had not responded by mail, yielding a total sample of 118 (58%) completed surveys. The verbal responses elicited by the phone contacts did not appear to be markedly different from responses to the mailed surveys.

Of the 106 departments offering their own beginning course, 73 (69%) used computers in this course. Thirty-three (31%) of the 106 departments did not use computers in the beginning course. However, 23 of these departments used computers in subsequent courses.

Twelve of the departments did not offer a beginning statistics course (students took the course in a mathematics or statistics department), but 10 of these used computer-assisted data analysis in upper level courses. In summary, 69% of the departments offering a beginning statistics course used computers in that course, and 90% of the departments sampled used computer-assisted data analysis in statistics or other courses.

Our tabulation of the variety of statistical software packages used showed that 59% of the 73 departments using computers in the beginning statistics course used SPSS (1999). These data are in keeping with the content of current statistics texts. For example, many texts devote at least part of their problem sets to computer solutions including SPSS, with some texts devoted exclusively to the SPSS package. Other texts include a large data set to accompany end-of-chapter computer exercises, whereas others may include an SPSS tutorial in an appendix.2 In contrast, the remaining 30 departments listed 14 different statistical packages or sources. Their choices ranged from the familiar Excel (1997), SAS (1999), and Minitab (1998) to Internet-based programs and those furnished by some statistics texts publishers.

Another item on the questionnaire concerned the instructional and practice location, with 26 (36%) of the 73 departments offering a regularly scheduled laboratory period for computer-assisted data analysis in their beginning statistics course. However, a variety of responses, including combinations of in-class, own time, and two or three or four time demonstration-practice sessions precluded any meaningful tabulation of the remaining 47 departments. Nevertheless, 36% of the departments did indicate a regularly scheduled laboratory period for data analysis.

There was minimal agreement on the last item on the questionnaire, which requested the title of any manual or workbook used with the computer-assisted data analysis. Forty-four of the 73 departments reported they did not use published materials but used their own handouts and worksheets or software tutorials. Another 25 did use a published user’s manual, but listed 18 different titles, with the most popular appearing only four times. The remaining 5 departments used their particular statistics text as a software guide.

Although the results of this survey were not totally unexpected, we were surprised by the finding that almost one third (31%) of the departments did not use computer-assisted data analysis in their beginning statistics course. It is not likely that budgetary considerations would be the sole explanation, given that 90% of the departments sampled used computer-assisted data analysis somewhere in their curriculum. Access to machines may be limited for instructors wanting computers for laboratory periods, but limited access in itself would not prevent instruction in computer-assisted data analysis.

2In an informal tabulation of the first author’s shelf of 21 statistics texts published since 1990, we found SPSS (1999) in one of these categories in 14 of the texts.
According to unsolicited comments from respondents, a more likely explanation for computer nonuse by a third of the departments is the hesitancy on the part of instructors to try to cover both statistical concepts and computer methodology in a single course. One of their concerns is that the average student who, with difficulty, is able to perform at a basic skills level in statistics alone, might be left behind in a combined course. For example, Rosen, Feeney, and Petty (1994) reported that students in one such combined course wanted more class time spent on hand calculations with a calculator. They also wanted more time devoted to class lectures on theory and derivations. Rosen et al. concluded that statistical packages should be used later in the curriculum rather than in introductory statistics. Earlier, Tromater (1985) described the benefits of designing a combined statistics and computer course to be taken after the introductory statistics course. In a recent interview in The Generalist’s Corner of Teaching of Psychology (Dillon, 1999), statistics textbook author Chris Spatz reported,

In addition, I understand from one textbook editor that many psychology professors are reluctant to have beginning students rely on computers because they think it will thwart students’ understanding of statistical tests. This concern about computers taking the place of understanding was also a concern of the APA Task Force on Statistical Inference. (p. 234)

In contrast, Ware and Chastain (1989) reported that students in computer-assisted statistics courses had similar course grades but more favorable attitudes toward statistics than those not using computer-assisted data analysis in the statistics course. They also noted that few studies on the efficacy of computer use in the teaching of the beginning course in statistics had been conducted. Their observation, unfortunately, holds true today.

Suggestions for future research are clear. If a substantial proportion of the third who do not offer computer-assisted data analysis with the beginning statistics course have legitimate pedagogical concerns for not doing so, then these concerns need to be examined. If the typical B and C students are achieving only marginal success in the combined computer-statistics course, a host of questions regarding teaching methods are raised. Are regularly scheduled laboratory periods essential? Should students generate their own data sets instead of relying on end-of-chapter exercises? How much calculator practice with traditional formulas is necessary prior to computer-assisted data analysis? Is it necessary at all?

These and many similar questions were undoubtedly faced in one form or another by the two thirds of the departments sampled who did offer computer-assisted data analysis in their beginning statistics course. If a series of empirical studies would show that these questions can be answered and that the average student can succeed, we would all like to know the details.

References


Notes

1. We thank David J. Pittenger, Mark Krejci, Susan Larson, and three anonymous reviewers for helpful comments on an earlier draft of this article.

2. Send correspondence to Albert E. Bartz, Department of Psychology, Concordia College, Moorhead, MN 56562; e-mail: bartz@cord.edu.
Correlational Analysis and Interpretation: Graphs Prevent Gaffes

Blaine F. Peden
University of Wisconsin–Eau Claire

In this article I describe an activity in which students use 4 data sets devised by Anscombe (1973) to enter data, compute Pearson rs, plot scatter graphs, and write results paragraphs. Although these data sets yield identical coefficients of correlation, \( r = .82 \), the scatter graphs reveal that only 1 data set is appropriate for Pearson correlational analysis and interpretation. Students who complete this assignment exercise their data entry, computational, graphical, and writing skills and also learn that graphs play an important role in good statistical analysis (i.e., graphs prevent gaffes). Teachers of psychology and authors of textbooks for statistics and research methods courses can adapt this exercise themselves or employ existing adaptations.

Every picture tells a story.
—Rod Stewart, 1971

Psychology students should learn about the concepts of correlation and regression in undergraduate statistics and research methods courses (Giesbrecht, Sell, Scialfa, Sandals, & Ehlers, 1997). Typically, such courses in psychology provide instruction about topics such as patterns in scatter graphs, correlation and linearity, least squares line, residual plots, outliers, influential points, and transformations to achieve linearity.

Despite the importance of these topics to students’ statistical and methodological understanding, only four articles in Teaching of Psychology have discussed teaching and learning about aspects of correlation. Duke (1978) indicated that students do not readily understand either size differences in correlation coefficients or the number of exceptions to the generalization implied by a particular value of \( r \). To remedy these problems, Duke devised tables that help students estimate the number of exceptions implied by the entire range of \( r \) values. Huck, Wright, and Park (1992) described a classroom demonstration that helps students understand three seemingly contradictory claims about the relation between variability and the size of a Pearson coefficient of correlation. Finally, both Goldstein and Strube (1995) and Mitchell and Jolley (1999) provided computer programs that teach students about the direction and size of \( r \) values and the influence of outliers on \( r \) values.

I created an activity in which students use four data sets devised by Anscombe (1973) to enter data, compute Pearson rs, plot scatter graphs, and write results paragraphs. Although these data sets yield identical coefficients of correlation \( (r = .82) \) the scatter graphs in Figure 1 reveal that only one data set is appropriate for a Pearson correlational analysis and interpretation. In Figure 1, Panel A depicts the mind’s eye image of \( r = .82 \), Panel B displays a curvilinear relation, and Panels C and D portray influential outliers.

![Figure 1. Scatter graphs of Anscombe's (1973) data sets. Panels A through D match data sets A through D in Table 1. Panel A depicts the mind's eye image of r = .82, Panel B displays a curvilinear relation, and Panels C and D portray influential outliers.](image-url)
illustrate influential outliers. Students who complete this assignment exercise their data entry, computational, graphical, and writing skills and also learn that graphs play an important role in good statistical analysis (Anscombe, 1973).

**Method**

**Participants**

Students were 28 women and 7 men enrolled in an undergraduate research methods course in the 1999 or 2000 spring semesters. All students had completed a prerequisite statistics course.

**Materials**

Table 1 presents the four bivariate data sets devised by Anscombe (1973). Typical correlation and regression analysis yields identical results for all the data sets: (a) $n = 11$, (b) mean of $x$s = 9.0, (c) mean of $y$s = 7.5, (d) equation of regression line: $Y = 3 + 0.5X$, (e) standard error of estimate of slope = 0.118, (f) $t = 4.24$, (g) sum of squares = 110.0, (h) residual sum of squares of $y = 13.75$, (h) $r = .82$, and (i) $r^2 = .67$. In sum, the same linear model describes Anscombe’s quartet until one examines the residuals.

**Procedure**

I distributed one page of instructions indicating that students should do four things to complete the assignment. First, the students selected a partner and read or reviewed discussions about correlation in their textbooks (Bordens & Abbott, 1999; Pavkov & Pierce, 2000; Stanovich, 1998). Second, the partners used SPSSolutions software (SPSS Inc., 1998) to create a data file, compute $r$ values, and graph each of the four data sets. Third, the partners wrote one paragraph per data set (a) indicating the $r$ value, (b) describing the relation between $x$ and $y$ in their own words, and (c) stating whether and why the data set was suitable for a Pearson correlational analysis and interpretation. Finally, the partners wrote a fifth paragraph that explained the lesson they learned from this assignment. One week later, I discussed the purpose and results of the exercise with the class prior to collecting the papers for grading.

**Results**

Students’ comments revealed that this assignment was straightforward, relatively easy to do, and effective in accomplishing its goals. Seventeen of the 18 partners reported the correct coefficients of correlation ($r = .82$) for all four data sets and presented appropriate scatter graphs. The one incorrect coefficient and inappropriate scatter graph could be traced to a data entry error. The majority (15 of 18) of partners accurately described the relation between $x$ and $y$ and also correctly explained whether and why each data set was or was not amenable to a correlational analysis and interpretation. The exceptions to this outcome typically resulted from failures to describe the relation between $x$ and $y$ at all. Finally, the majority (16 of 18) of partners adequately expressed a version of the lesson to be learned. For example, one pair of students indicated:

By simply comparing the scores of the four different data sets, one would conclude that the distributions were very similar in nature. However when the distributions are presented visually as in the scatter plots we can see that there are many different factors affecting the Pearson score obtained.

A second pair of students stated:

This exercise shows us that correlation coefficients may be the same, but may have very different scatter plots. We need to look at both the correlation coefficient and the graph to determine the meaning of the correlation coefficient.

**Discussion**

Instructors who use this exercise likely will obtain comparable results with their students. Their students will exercise their data entry, computational, graphical, and writing skills

<table>
<thead>
<tr>
<th>Case</th>
<th>$Ax$</th>
<th>$Ay$</th>
<th>$Bx$</th>
<th>$By$</th>
<th>$Cx$</th>
<th>$Cy$</th>
<th>$Dx$</th>
<th>$Dy$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10.0</td>
<td>8.04</td>
<td>10.0</td>
<td>9.14</td>
<td>10.0</td>
<td>7.46</td>
<td>8.0</td>
<td>6.58</td>
</tr>
<tr>
<td>2</td>
<td>8.0</td>
<td>6.95</td>
<td>8.0</td>
<td>8.14</td>
<td>8.0</td>
<td>6.77</td>
<td>8.0</td>
<td>5.76</td>
</tr>
<tr>
<td>3</td>
<td>13.0</td>
<td>7.58</td>
<td>13.0</td>
<td>8.74</td>
<td>13.0</td>
<td>12.74</td>
<td>8.0</td>
<td>7.71</td>
</tr>
<tr>
<td>4</td>
<td>9.0</td>
<td>8.81</td>
<td>9.0</td>
<td>8.77</td>
<td>9.0</td>
<td>7.11</td>
<td>8.0</td>
<td>8.84</td>
</tr>
<tr>
<td>5</td>
<td>11.0</td>
<td>8.33</td>
<td>11.0</td>
<td>9.26</td>
<td>11.0</td>
<td>7.81</td>
<td>8.0</td>
<td>8.47</td>
</tr>
<tr>
<td>6</td>
<td>14.0</td>
<td>9.96</td>
<td>14.0</td>
<td>8.10</td>
<td>14.0</td>
<td>8.84</td>
<td>8.0</td>
<td>7.04</td>
</tr>
<tr>
<td>7</td>
<td>6.0</td>
<td>7.24</td>
<td>6.0</td>
<td>6.13</td>
<td>6.0</td>
<td>6.08</td>
<td>8.0</td>
<td>5.25</td>
</tr>
<tr>
<td>8</td>
<td>4.0</td>
<td>4.26</td>
<td>4.0</td>
<td>3.10</td>
<td>4.0</td>
<td>5.39</td>
<td>19.0</td>
<td>12.50</td>
</tr>
<tr>
<td>9</td>
<td>12.0</td>
<td>10.84</td>
<td>12.0</td>
<td>9.13</td>
<td>12.0</td>
<td>8.15</td>
<td>8.0</td>
<td>5.56</td>
</tr>
<tr>
<td>10</td>
<td>7.0</td>
<td>4.82</td>
<td>7.0</td>
<td>7.26</td>
<td>7.0</td>
<td>6.42</td>
<td>8.0</td>
<td>7.91</td>
</tr>
<tr>
<td>11</td>
<td>5.0</td>
<td>5.68</td>
<td>5.0</td>
<td>4.74</td>
<td>5.0</td>
<td>5.73</td>
<td>8.0</td>
<td>6.89</td>
</tr>
</tbody>
</table>

Note. Attentive students can hasten data entry and minimize errors by recognizing that the $Ax$, $Bx$, and $Cx$ values are identical; less attentive students simply get more practice entering data.
and also learn what Anscombe (1973) meant when he said "graphs are essential to good statistical analysis" (p. 17) and why Tufte (1983) contended scatter "graphics can be more precise and revealing than conventional statistical computations" (p. 13). More colloquially, this exercise teaches students that graphs prevent gaffes when analyzing and interpreting correlations. The results from this activity also support the contention that psychology instructors should provide more explicit instruction regarding the role of graphs in the research process (Peden & Hausmann, 2000; Pittenger, 1995).

Students who complete this assignment learn some of the same things as students who use the computer programs devised by Goldstein and Strube (1995) or Mitchell and Jolley (1999). Nonetheless, one advantage of this activity is that students actively engage in several steps of the research process. That is, students must enter data, compute Pearson rs, create scatter graphs, and write results in American Psychological Association-style paragraphs. Perhaps the most effective instructional strategy would be one in which students learn about correlation from their instructors and text, work with the computer programs, and then complete this exercise on their own.

Teachers of psychology can adapt this exercise or Anscombe's (1973) data sets for use in a computer laboratory or traditional classroom. For example, my colleague William Frankenberger (personal communication, June 12, 2000) met his elementary statistics class in a computer lab. After the students entered each data set, computed summary statistics for a simple regression analysis, and obtained a scatter graph with a regression line, the class discussed various issues. The discussion of Data Set A established the standard analysis and interpretation of Pearson correlation coefficients and simple regression analysis and also provided a framework for examining the other three data sets. The discussion of Data Set B concerned violation of the rectilinearity assumption and how predictive ability could be improved by use of curvilinear regression. The discussions of Data Sets C and D focused on how inspection of the scatter graph reveals the influence of outliers that undermine the ability to predict Y given a value of X.

A statistics or methods instructor in a traditional classroom could distribute one of four worksheets, each displaying Anscombe's (1973) original data sets or the modifications. That is, students must enter data, compute Pearson rs, create scatter graphs, and write results in American Psychological Association-style paragraphs. Perhaps the most effective instructional strategy would be one in which students learn about correlation from their instructors and text, work with the computer programs, and then complete this exercise on their own.

Teachers of psychology can also employ existing adaptations of Anscombe's (1973) data sets. For example, Lorenz (1987) expressed a concern about the tendency of introductory statistics courses in the social sciences to underplay the idea that some outliers have more influence than other outliers. To remedy this situation, he devised three different twelfth points that could be added to Data Set A in Table 1. Lorenz also taught his students to evaluate residuals by examining (a) the size of the residual; (b) its leverage; and (c) Cook's distance, a summary measure of influence. In addition, Bremner (1999) posted modifications to the Anscombe regression data sets and also added a fifth data set. Bremner's data sets and exercises also appeared in a section on regression and correlation in a volume by Bassett et al. (1986).

Finally, authors of statistics and research methods textbooks could discuss the lessons to be learned from Anscombe's (1973) original data sets or the modifications. Alternatively, authors could incorporate this exercise or adaptations as homework exercises.

References


Notes

1. A preliminary report about this exercise was given at the Midwestern Psychological Association in Chicago, May 2000.

2. I thank Allen Keniston, Ken McIntire, and Doug Woody for helpful comments on this article.

3. Send correspondence to Blaine F. Peden, Department of Psychology, University of Wisconsin, Eau Claire, WI 54702–404; e-mail: pedenbf@uwec.edu.
Microsoft Excel™ as a Tool for Teaching Basic Statistics
C. Bruce Warner and Anita M. Meehan
Teaching of Psychology 2001 28: 295
DOI: 10.1207/S15328023TOP2804_11

The online version of this article can be found at:
http://top.sagepub.com/content/28/4/295

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Society for the Teaching of Psychology

Additional services and information for Teaching of Psychology can be found at:

Email Alerts: http://top.sagepub.com/cgi/alerts

Subscriptions: http://top.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Oct 1, 2001

What is This?
Many instructors use specialized statistics packages, such as SPSS or SAS/STAT in teaching quantitative and research methods courses. As an alternative to the specialized packages, Microsoft Excel™ offers charting capabilities and the Analysis ToolPak, which provides a set of data analysis tools beyond those normally found in spreadsheets. With Analysis ToolPak, Excel resembles what one would expect from a specialized low-level statistics program. We have successfully integrated Excel into our introductory statistics courses and developed a tutorial manual to guide students through various statistical procedures in Excel. We present student ratings of an intermediate form of the manual and its exercises and discuss the advantages and disadvantages of using Excel rather than a specialized statistical package for teaching basic statistics.

Similar to many instructors in the social and behavioral sciences, for years we relied primarily on specialized statistical software such as SPSS in our statistics and research methods courses. As computer technology has become ubiquitous and spreadsheet programs more popular, we wondered if a spreadsheet application was not a better medium for a first course in statistics. Keeping in mind that most students would be expected to use spreadsheets in their work places (Davis, 1997), we began to use Microsoft Excel™, a popular spreadsheet software package, in our introductory statistics courses.

Why Use a Spreadsheet Application Such as Excel?

One reason we embraced a spreadsheet application is that students need data analysis and computer skills that they can generalize and transfer to many contexts. Students are likely to work in environments in which general software skills, such as use of spreadsheet applications, are expected (Olsten Corporation, 1996). A recent survey found that prospective employers ranked spreadsheet skills second, behind word-processing skills, as highly desirable and expected skills for new hires (Davis, 1997). Moreover, 86% of the employers expected new hires to go beyond spreadsheet basics and be able to perform detailed analyses. Davis also found that employers highly valued the ability to create presentations and graphs.

Another reason we decided to teach with Excel is that we believe its simplicity facilitates student learning (Henle, 1995). Students can concentrate more on statistics and less on the mechanics of the software. Data entry is often easier in spreadsheet applications than in specialized statistical programs such as SPSS. Moreover, many students already have some familiarity with spreadsheets, thereby reducing the time instructors must devote to software training. In our experience, the sophisticated output generated by advanced statistical packages can be confusing for students. The output generated by Excel often matches examples presented in introductory statistics textbook exercises or in classroom instruction.

Another advantage of Excel is that it simplifies the electronic exchange of assignments. When a student turns in an assignment on disk or via e-mail, the instructor can open the file, examine the contents, and attach comments directly to the file using the annotation tools. For instance, when a student turns in an assignment in which he or she has performed the wrong test, the instructor can insert a text balloon that points directly to the error and offers corrective advice. When we grade assignments, we save a copy of each text balloon into an open workbook, so that we can later paste the balloon into the workbook of another student who has made a similar mistake.

A final advantage of using Excel is economic. Students and universities already own or have access to computers with spreadsheet software, whereas specialized programs are an added expense. The extra cost also means that students have limited access to the specialized statistical software, as it is only available in selected campus locations.

Why Choose Excel?

We were familiar with Excel because it is the spreadsheet standard for our campus. Still, it is a natural choice for a statistics course, as it performs most basic analyses, creates easily customizable graphs using its Chart Wizard feature, and is the market leader in its product category.

Similar to all spreadsheet applications, Excel has the capability to solve both simple and complex equations. Excel's Paste Function (known as Function Wizard in versions earlier than Excel 95) simplifies the process by taking you through the steps necessary to paste statistical calculations into your spreadsheet. For example, you can use Paste Function to find common statistics such as mean, median, mode, standard deviation, variance, and correlation.

Excel also contains a less commonly known feature: Analysis ToolPak. Once you have performed the simple steps that install this add-in as a working part of your Excel application, you have an array of tools for performing descriptive statistics, t tests, ANOVA (one-way, one-way repeated measures, and two-factor), correlation, and regression. It also adds statistical and graphing capabilities that are unavailable in the basic Excel application. To illustrate the relative simplicity of
the Analysis ToolPak, the top panel of Figure 1 contains an example of the dialog box that appears when you select an independent-samples t-test analysis. In the dialog box, “input variable range” refers to the spreadsheet location for the data from each group. In our example, we had data in two spreadsheet columns representing the hypothetical pain-pill consumption of a meditation group (A1–A13) versus a control group (B1–B13). From the dialog box, you can also choose to include column labels with the data being input. Labels help to clarify the output from the procedure. The Excel output appears in the bottom panel of Figure 1.

Additional resources exist to support instruction with Excel, including various statistical add-ins and simulations for Excel written in Visual Basic. Internet discussion groups have evolved around the topic of using spreadsheets in mathematics and science education, including teaching statistics. Good starting points for Excel resources on the Web are the Association of Statistics Specialists Using Microsoft Excel (ASSUME; http://www.jiscmail.ac.uk/lists/assume.html) and Spreadsheets in Education (http://sunsite.univie.ac.at/Spreadsite/).

**Student Evaluation of Excel**

Beginning around 1996, we incorporated a few Excel assignments into our introductory statistics course and developed accompanying handouts. For the Fall 1998 semester, we expanded our use of Excel and developed an instruction manual for students. The manual explained how to use the Paste Function, Analysis ToolPak, and Chart Wizard to perform many of the basic procedures covered in an introductory statistics course. The manual included screen shots of Excel (e.g., those in Figure 1), sample problems with associated statistical output, and homework assignments. An example of a homework assignment follows:

Make up a data set that is appropriate for an independent-samples t-test and label your two groups. Use 15 scores per group. Do the analysis and save the output as Sheet 4. Beneath the output generated by Excel, type in your interpretation of the results. This must include: the value of the observed t and its actual p value, the critical t, the means for each group, and a decision about the null hypothesis. Use alpha = .05, two-tailed. See the model provided for interpreting results under Step 4 of the section on computing an independent-samples t test using Excel.

We required students to generate simulated data to ensure that each submission was original and to provide students with practice in generating simple data sets appropriate for analysis. By examining the data and interpretations students had generated, we could often discover and correct misconceptions that students had about various statistical procedures. For instance, some students persisted in generating scores from two imaginary independent samples when performing a related-samples t test.

Students completed five required Excel assignments during the semester. In addition, almost all students chose to turn in at least one bonus assignment, with the majority completing three. To improve our teaching of statistics as well as our instruction manual, we decided to conduct a formal evaluation of student reactions to Excel.

**Method**

**Participants**

We surveyed 49 of the 52 students enrolled in our Fall 1998 Introductory Statistics classes who had used the manual and completed the Excel assignments during the semester. Participation in the survey was voluntary and anonymous.

**Measures**

We asked the students to express their opinions about the homework assignments and the course manual using a four-part questionnaire. In Parts A through C, students rated the assignments and manual. Students responded to semantic differential descriptions of the assignments (Part A) and the course manual (Part C) on a 5-point scale, ranging from 1 (complete agreement with the description on the left) to 5 (complete agreement with the description on the right). In Part B, participants responded to course goal-oriented statements about the Excel assignments. Ratings on Part B were also based on a 5-point scale ranging from 1 (strongly disagree) to 5 (strongly agree). The complete set of items and the mean responses appear in Table 1. Part D prompted students to provide free-form comments and suggestions.
Results and Discussion

Table 1 presents the results of the student ratings of the Excel homework assignments. Students indicated that the assignments were useful, not too much work, and required some degree of creativity and critical thinking. They expressed agreement that the assignments promoted the development of useful skills. Although students agreed that the assignments improved their knowledge of statistical concepts, they rated the assignments more highly in terms of improving their computer skills. Student frustration with the assignments was moderate, which we believe is a fairly reasonable level given the typical frustrations associated with any statistics assignment, plus the added headaches associated with using any software application. In many instances, frustrations were not specifically related to Excel but were tied to general computer issues such as getting access to campus labs or forgetting to save a homework file to disk. Reactions to the instruction manual were also generally positive, as shown in Table 1. Students indicated the appropriateness of both the length and comprehensiveness of the manual. They found the manual helpful for completing assignments and generally found it useful and user friendly.

<table>
<thead>
<tr>
<th>Evaluation Item</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assignments—Part A: Semantic differential items</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extremely bad–extremely good</td>
<td>3.59</td>
<td>0.61</td>
</tr>
<tr>
<td>Boring–interesting</td>
<td>3.40</td>
<td>0.89</td>
</tr>
<tr>
<td>Difficult–easy</td>
<td>3.51</td>
<td>0.98</td>
</tr>
<tr>
<td>Unpleasant–pleasant</td>
<td>3.39</td>
<td>1.02</td>
</tr>
<tr>
<td>Not frustrating–frustrating</td>
<td>3.06</td>
<td>1.22</td>
</tr>
<tr>
<td>Imposing–friendly</td>
<td>3.65</td>
<td>0.81</td>
</tr>
<tr>
<td>Confusing–clear</td>
<td>3.43</td>
<td>1.00</td>
</tr>
<tr>
<td>Too much work–not too much work</td>
<td>3.84</td>
<td>0.85</td>
</tr>
<tr>
<td>Not useful–useful</td>
<td>3.86</td>
<td>1.04</td>
</tr>
<tr>
<td>Assignments—Part B: Likert-scale items</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Excel assignments helped me understand statistical concepts better</td>
<td>3.37</td>
<td>1.04</td>
</tr>
<tr>
<td>The Excel assignments improved my computer skills</td>
<td>3.94</td>
<td>0.99</td>
</tr>
<tr>
<td>The Excel assignments allowed me to be creative</td>
<td>3.65</td>
<td>0.90</td>
</tr>
<tr>
<td>The Excel assignments required critical thinking on my part</td>
<td>3.69</td>
<td>0.96</td>
</tr>
<tr>
<td>The Excel assignments gave me skills I can use</td>
<td>3.90</td>
<td>1.02</td>
</tr>
<tr>
<td>Excel manual—Part C: Semantic differential items</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extremely bad–extremely good</td>
<td>3.96</td>
<td>0.64</td>
</tr>
<tr>
<td>Not helpful–helpful</td>
<td>4.20</td>
<td>0.89</td>
</tr>
<tr>
<td>Hard to follow–easy to follow</td>
<td>3.69</td>
<td>1.07</td>
</tr>
<tr>
<td>Not detailed enough–too detailed</td>
<td>3.04</td>
<td>0.71</td>
</tr>
<tr>
<td>Too short–too long</td>
<td>2.98</td>
<td>0.32</td>
</tr>
<tr>
<td>Imposing–friendly</td>
<td>3.84</td>
<td>0.85</td>
</tr>
<tr>
<td>Confusing–clear</td>
<td>3.49</td>
<td>1.00</td>
</tr>
<tr>
<td>Not useful–useful</td>
<td>4.25</td>
<td>0.80</td>
</tr>
</tbody>
</table>

Note. N = 49. For the 5-point ratings on parts A and C, scores ranged from 1 (complete agreement with the description on the left) to 5 (complete agreement with the description on the right). The poles for several items on the survey were reversed to reduce participant response bias. In Part B, scores ranged from 1 (strongly disagree) to 5 (strongly agree).

Considerations for Using Excel to Teach Statistics

Deciding whether to use Excel rather than a specialized statistical package not only requires a careful comparison of the specific merits of each software package, but also demands a methodical examination of student needs, course and department goals, and instructor expectations. Excel makes sense when time for teaching software is limited or access to specialized software is limited, when course goals include the development of general computer skills, when most students will enter the job market directly and not pursue graduate training, or when students come from a variety of departments that emphasize different statistical packages in training their majors. An instructor considering the adoption of Excel should also consider whether students would have an opportunity to learn a specialized statistical package in another course. In our opinion, Excel training should not constitute a student’s sole exposure to computerized data analysis. Although we may use Excel in introductory statistics, we use both Excel and SPSS in our methods courses.

Excel simply cannot supplant specialized packages for two reasons. First, the standard set of statistical procedures for Excel is quite limited compared to packages such as SPSS. For instance, Excel can perform only three ANOVA procedures: (a) one-way ANOVA for independent samples, (b) one-way repeated-measures ANOVA, and (c) two-way ANOVA for independent samples. Excel’s ANOVA procedures will not perform planned comparisons or post-hoc tests, and the two-way ANOVA requires equal Ns in all cells. Second, the standard statistical procedures included in Excel are not as robust as procedures in specialized statistical packages (e.g., McCullough & Wilson, 1999; Sawitzki, 1994). Based on Excel’s performance on a variety of data sets, McCullough and Wilson, and Sawitzki, recommended that Excel not be
used for scientific data analysis. They cited weakness in Excel's implementation of several algorithms, which will lead to serious errors on large data sets, especially those containing large values. McCullough and Wilson also observed that Microsoft does not correct flawed algorithms, even though the flaws have been exposed in earlier versions of Excel.

Because Excel has a limited variety of procedures and because it may produce inaccurate results on large data sets, it is simply not a good choice for advanced courses or general scientific data analysis. Realistically, though, most introductory courses will not cover many procedures beyond those included in Excel's Analysis ToolPak, and sample data sets will be no larger than those that can be computed conveniently by hand. Excel's performance on small data sets, in our experience, is generally very acceptable (cf. McCullough & Wilson, 1999). What Excel does offer for introductory courses is simplicity of both input and output, which Henle (1995) described as transparency, meaning that students can easily see the relation between the numbers and the results of operations performed on those numbers. Also, add-ins are available for Excel that may overcome limitations of both flexibility and accuracy. A list of available add-ins can be found at the ASSUME Web site (http://www.jiscmail.ac.uk/lists/assume.html).

Regardless of whether the instructor chooses Excel or a specialized package, students expect training on the software and expect guidance in completing computer assignments. In a recent survey of over 1,000 Cornell University undergraduates, students rated credit-bearing classes and peer support as the most effective means of acquiring computer skills (Davis, 1999). A small student focus group also expressed how frustrating and stressful it is when professors assume students possess certain software skills yet fail to provide students with support or training (Davis, 1999). At the same time, the focus group also indicated that students do not want professors to devote significant class time to teaching software at the expense of course content. Based on our experience, and consistent with the student views expressed in Davis (1999), we believe instructors need some sort of student guide (e.g., Berk & Carey, 2000; Dretzke & Heilman, 1998; Meehan & Warner, 2000) to help students complete assignments efficiently, effectively, and with minimal frustration.

Conclusions

We believe that Excel is a useful tool for teaching elementary statistics. Business schools have been at the forefront of using spreadsheet applications in educating their students, and many resources for business statistics exist (e.g., Levine, Berenson, & Stephan, 1999; Middleton, 2000; Neufeld, 1998; Pelosi & Sandifer, 1999), but spreadsheet-oriented statistics guides for broader audiences have also been appearing (e.g., Berk & Carey, 2000; Dretzke & Heilman, 1998; Meehan & Warner, 2000). Individual instructors will need to weigh the costs and benefits of using a spreadsheet program versus a specialized statistical package, but we believe a spreadsheet, such as Excel, will prove more attractive in many situations.

References


Notes

1. Support for the initial development of an Excel instruction manual was provided by a grant from the Kutztown University Professional Development Committee.

2. A version of this article was presented at the meeting of the American Psychological Society, Denver, CO, June, 1999.

3. We thank Megan Mumma, who student tested our tutorials. We also thank David Pittenger, Randolph A. Smith, and three anonymous reviewers for their helpful comments on earlier drafts of this article.

4. Send correspondence to C. Bruce Warner, Department of Psychology, Kutztown University, Kutztown, PA 19530; e-mail: warner@kutztown.edu.
A Review of ESTAT: An Innovative Program for Teaching Statistics
Michael A. Britt, John Sellinger and Lee M. Stillerman
Teaching of Psychology 2002 29: 73
DOI: 10.1207/S15328023TOP2901_13

The online version of this article can be found at:
http://top.sagepub.com/content/29/1/73

Published by:
http://www.sagepublications.com

On behalf of:
Society for the Teaching of Psychology

Additional services and information for Teaching of Psychology can be found at:

Email Alerts: http://top.sagepub.com/cgi/alerts
Subscriptions: http://top.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Jan 1, 2002

What is This?
A Review of ESTAT: An Innovative Program for Teaching Statistics

Michael A. Britt
John Sellinger
Lee M. Stillerman
Marist College

We review the benefits of using an interactive, graphics-based program called Estimating Statistics (ESTAT) as a teaching tool for statistics with psychology students. ESTAT incorporates two features not found in other programs: eyeball estimation and instantaneous statistical display. We describe an overview of the statistical concepts covered by the program, its unique pedagogical features, and the exercises used to address these concepts. The majority of our students recommended that we use this program in the future. In addition, we include a critique of the program.

Psychology students sometimes approach statistical analysis with apprehension. Fortunately, many software packages and Web-based tutorials make statistical analysis more easily understandable for students. One such program is Estimating Statistics (ESTAT) for Comprehending Behavioral Statistics. This program uses graphical and computational components, both of which are interactive by design. The graphical components of ESTAT, including histograms, scatterplots, and tables, serve to make important statistical concepts visible to the student, thus diminishing the abstract nature of statistics that sometimes elicits feelings of anxiety. In addition, many of the exercises in ESTAT require the student to become engaged in a nontargeting way by guessing the magnitude of a statistic or grabbing points or lines on the computer screen and moving them around scatterplots. Overall, these activities are student-friendly, which helps in making the process of statistical analysis less threatening and more educational.

Some attempts at using computers as teaching tools in the classroom have been very successful, as demonstrated by Erwin and Rieppi (1999) in their comparison of multimedia and traditional approaches in teaching undergraduate psychology courses. These researchers wanted to determine whether students taught in a multimedia classroom would have higher scores on several psychology tests than those taught in a traditional lecture classroom. They found that students taught in a multimedia classroom did score significantly higher on their final exam than those taught in the traditional classroom.

Computer technology has also demonstrated its usefulness in other academic disciplines. Misale, Gillette, and delMas (1996) found computer technology to be an effective teaching tool in the disciplines of economics and psychology. Using the skill of decision making, these researchers utilized computer programs for a variety of tasks, including data entry and presentation, complex calculations, and graphical output. Decision making was also explored using group discussion, debate, and exercises. Feedback indicated that students felt highly involved in the classroom activities and praised the use of the computer as an alternative way of addressing the topic. Students gave favorable ratings to the interactive nature of the program and the use of the graphical output obtained from the program. Such interactive exercises and graphical representations of data are two of the defining features of the ESTAT program.

Program Description

ESTAT is a computer program that provides activities to help students learn many different statistical concepts, from descriptive statistics to comprehensive inferential statistical tests. Among the statistical concepts addressed are the mean, standard deviation, area under the normal curve, correlation, confidence intervals, t tests, ANOVA, and regression. Each of these concepts is addressed through activities with varying levels of difficulty, from simply observing a graphical representation of data and eyeball estimating the mean of the distribution, to creating data for several groups and entering this data into the program’s spreadsheet for instant analysis.

At the basic level, ESTAT provides an interactive review of the mean and the standard deviation. One feature provided by ESTAT is the unique way it presents data in a histogram. Students “eyeball estimate” the mean of the data by dragging a vertical arrow along the x axis to the point where they believe the mean of the distribution lies. Once the mean is established, the student then draws a horizontal line along the x axis to a width that he or she believes represents the size of the standard deviation of the distribution. Once these lines are in place, the student guesses the precise values of the mean and standard deviation. The program compares student answers to the actual values and calculates an eyeball estimation score. This score tells students how close their eyeball estimate is to the actual statistic.

The exercise for the t test presents students with two histograms representing data from two samples. Student have to guess the mean of each distribution and then carefully examine the variances of each distribution. They guess at the standard error of the differences and make guesses regarding how large the observed value of t is likely to be and whether there
will be a statistically significant difference between the two distributions.

Another unique activity addresses the ANOVA procedure. ESTAT provides what it terms paths that assist the student in addressing the concept at varying levels of difficulty. The easier of the ANOVA paths provides the student with an ANOVA table that contains only two values. From these given values, the student must infer the other values or values to input. This activity gives the student practice in exercising the basic relations of the values in an ANOVA table. One of the more difficult paths presents the student with four columns of data, accompanied by a corresponding completed ANOVA table. On another side of the screen are four additional columns of data, accompanied by a blank ANOVA table. This second set of data is similar to the first data set, except that one column of data is changed in some way (e.g., each value is increased by five points). The challenge for the student is to answer several questions regarding the makeup of the new ANOVA table as it relates to the original ANOVA table. For example, the student must respond to questions such as “Will the spread of the sample means be larger, smaller, or the same?” and “Will this change in the data cause the within- and between-group variance to increase, decrease, or remain the same?” The student must realize that increasing each value in one column of data will not affect the within-groups variance, but will increase the between-groups variance and will thus also increase the F ratio. To answer such questions, students are challenged to exercise a thorough understanding of the inner workings of the ANOVA procedure.

In the regression section, students see a scatterplot of data and a flat line of “best fit.” Students must grab each end of this line with the cursor and adjust the line to where they believe it will be the line of best fit to the data. They then guess the exact value of the intercept and the slope. Students input their guess, and ESTAT graphs the actual line of best fit. It also provides the actual value of the intercept, the slope, and the standard error of the estimate.

In addition to the eyeball-estimation activities, ESTAT also provides a unique computational utility called datagen that allows the student to enter data into a spreadsheet for instant analysis. The spreadsheet is similar to other larger statistical software packages. What distinguishes this spreadsheet is the fact that the analysis is immediately calculated while the student enters data. For example, as the student enters data in the first column of the spreadsheet, ESTAT calculates basic statistics such as the sample size, mean, standard deviation, and the standard error. These numbers change automatically as students continue to enter data. If the student enters data in a second column, datagen automatically calculates statistics such as the between- and within-groups t test. Once both columns have an identical number of data points, ESTAT automatically calculates the Pearson correlation coefficient. Finally, if data are entered into the third column and beyond, ESTAT automatically creates an ANOVA summary table and instantly displays it, along with a repeated-measures ANOVA if the sample sizes happen to be equal. The benefit of such real-time analysis is that students can change any of the values in the spreadsheet and instantaneously see how that change alters various statistical values. This approach gives them a better understanding of how these statistics operate.

Some of the activities we have described can be found on the Internet. However, having such a powerful program housed on one disk eliminates many of the headaches that the Internet often causes. For one, having this program run from one’s own computer eliminates the wait time often experienced when using the Internet. One can also rely on this program to be operational at any time, because the need for a live connection to the Internet is not required. Finally, the allowance of quick, real-time calculations by this program make it a truly exceptional educational experience for students, as they have the opportunity to see not only how statistical values are created and altered, but also what factors affect the values of a particular statistical result.

**Student Evaluations**

We observed how students interacted with the program when using it as part of a statistics review in a psychology research methodology course. In particular, we saw how the program helped engage students in statistics when we combined exercises from the program with a competitive game that we created. This game involved a series of activities with varying levels of difficulty, each requiring students, working in teams of four, to make eyeball estimates of data and to create data to meet certain statistical criteria. For example, teams examined a histogram created by ESTAT and then had to guess both its mean and standard deviation. The program quickly provided the correct answer and the team with the closest guess received a point. Also, teams observed a scatter diagram presented by ESTAT and then ventured a guess as to the magnitude of the correlation. The team with the closest guess received a point. Another game consisted of providing students with fictional midterm grades and then requiring each group to create final exam grades that would correlate at the .75 level (or final course grades that would correlate negatively with number of absences). Once students created the data, they used ESTAT’s datagen function to quickly calculate the results and determine which group was closest to the desired correlation. We examined each groups’ data and discussed why their guess was close or far away from the desired result. We also changed a few numbers in the data set to discover whether this change would bring us closer or farther away from the desired correlation. ESTAT quickly and automatically recalculates the correlation when the new data are entered. These activities, combined with the competition for points, made the review of these statistical concepts both engaging and enlightening.

As a follow-up to this activity, we obtained feedback from students to gain their opinion on the utility of this program. We collected data on a variety of issues, including student interest in the program, the degree to which students found the program to be engaging, the extent to which students believed the program helped them to better understand statistics, and whether students would recommend the use of this program in future classes.

The 5-point Likert scale used for the survey items ranged from 1 (strongly disagree) to 5 (strongly agree). In response to a question about interest in the program, the mean response was 4.10 (SD = .72), with most students (69.2%) indicating that
they agreed with the statement regarding interest. In addition, students found the program to be engaging, with a mean response of 3.90 ($SD = .78$) and 59.3% of respondents indicating that they agreed that the program was engaging. With students finding the program to be both interesting and engaging, it is no surprise that they reported the program to be useful in helping them to better understand statistics, as indicated by a mean response of 3.60 ($SD = .84$). Likewise, a mean response of 3.90 ($SD = .70$) demonstrates that students thought the program was a good review of statistics. As a result, 86.8% of respondents either agreed or strongly agreed that they would recommend the use of this program in future classes, as evidenced by a mean response of 4.00 ($SD = .73$).

Criticisms

In addition to mentioning the advantages of using this program, a thorough review would not be complete without mention of some of the shortcomings. We believe that a few statistical concepts could be added to make this program even more effective as a teaching tool. First, ESTAT does not include a post hoc test for use with the ANOVA. The inclusion of a Tukey Honestly Significant Difference test or a Neuman–Keuls could increase the utility of this path of the program. In addition, ESTAT does not include any nonparametric tests such as the chi-square. Inclusion of these concepts in ESTAT would increase the utility of this already exceptional program. ESTAT is available for Windows (3.1 or higher), Macintosh (System 6 or higher), and DOS machines. The DOS version contains most of the features of the other versions, but the interface is not graphical, and so should be used only when Windows or Macintosh machines are not available.

Conclusions

ESTAT is an innovative program that addresses important statistical concepts in a way that is user friendly. With the ability to employ both numerical and graphical representations of data in an interactive fashion, this program distinguishes itself as one of the more innovative software packages for teaching statistics and the only one that uses eyeball estimation as a way of cementing conceptual statistical knowledge. In addition, the comprehensive nature of the review of statistics offered by ESTAT makes it a program that is useful and challenging for individuals with varying degrees of statistical knowledge. With the current literature citing the benefits of computer technology in the classroom, this inexpensive program provides the student with the necessary tools to make such classroom integration a reality. In addition, this innovative program provides the student with the peace of mind that comes from knowing that the technology is always at their fingertips and not on remote servers that can often be slow and unreliable. Whether the student is looking for a simple review of statistical concepts or a comprehensive understanding of how data input can alter statistical results, ESTAT is an excellent choice.

References


Notes

1. ESTAT is available from Wadsworth Publishing at www.wadsworth.com.
2. Send correspondence and requests for more information about the competitive game discussed in this article to Michael A. Britt, Department of Psychology, Marist College, Poughkeepsie, NY 12601; e-mail: mbritt@activelearningtech.com.
Evaluation of an Interactive Tutorial for Teaching Hypothesis Testing Concepts
Christopher L. Aberson, Dale E. Berger, Michael R. Healy and Victoria L. Romero
Teaching of Psychology 2003 30: 75
DOI: 10.1207/S15328023TOP3001_12

The online version of this article can be found at:
http://top.sagepub.com/content/30/1/75

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Society for the Teaching of Psychology

Additional services and information for Teaching of Psychology can be found at:

Email Alerts: http://top.sagepub.com/cgi/alerts
Subscriptions: http://top.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Jan 1, 2003
What is This?
In this article, we describe and evaluate a Web-based interactive tutorial used to present hypothesis testing concepts. The tutorial includes multiple-choice questions with feedback, an interactive applet that allows students to draw samples and evaluate null hypotheses, and follow-up questions suitable for grading. Students either used the interactive tutorial \((n = 15)\) or completed a standard laboratory assignment \((n = 10)\) covering the same topics. Students who used the tutorial performed better \((p = .06)\) on a quiz than students who completed the standard laboratory, supporting the effectiveness of this freely available online tutorial. A second group of students \((n = 112)\) who did not participate in the assessment overwhelmingly rated the tutorial as easy to use, clear, and useful.

Null hypothesis significance testing (NHST) procedures are a primary focus of introductory statistics courses (Friedrich, Buday, & Kerr, 2000). Recent criticisms of NHST focus on misunderstandings and misuses of hypothesis testing (e.g., Cohen, 1994; Nickerson, 2000). Nonetheless, hypothesis testing remains central to psychology students’ abilities to understand research reports and conduct independent research. As such, it is essential that students gain a detailed understanding of hypothesis testing. We believe that many statistics courses emphasize rote learning of hypothesis testing, focusing on mechanical approaches to drawing statistical conclusions. Students learn to reject the null hypothesis if a computed value is larger than a comparison value. Although a mechanistic approach allows students to produce correct answers in simple situations, students who learn by rote are less likely to develop a deeper understanding of topics (Lovett & Greenhouse, 2000). Consequently, misunderstandings may occur regarding interpretation of statistical significance.

The Web Interface for Statistics Education (WISE) project provides instruction that addresses these shortcomings. As part of the WISE project, we created several Web-based tutorials that require only a JAVA-enabled browser. In this article, we present an interactive tutorial to assist students in learning about hypothesis testing with the \(z\) distribution. We use a normal distribution approach, as it is our impression that many introductory statistics courses use \(z\) as an introduction to hypothesis testing. The tutorial (found at http://wise.cgu.edu under “tutorials”) consists of a paper-based assignment that guides students’ use of an interactive applet, follow-up questions appropriate for grading or discussion, and on-screen multiple-choice questions that allow students to gauge understanding as they progress through the tutorial. This assignment assumes knowledge of \(z\) and normal distribution probabilities as well as some in-class introduction to hypothesis testing.

The tutorial begins with a description of a research scenario. The task is to investigate the effectiveness of three training programs. Students examine the mean and standard deviation of scores for a population of students who took a standardized test but did not take part in any training course (null population) and the means for populations of students who completed one of three training programs. One program is very effective (i.e., program mean is much larger than the mean for students with no training), the second is moderately effective, and the third is slightly effective. The effectiveness of the three programs corresponds to Cohen’s large, medium, and small effect sizes when compared to the null population (Cohen, 1988).

Next, the student answers a series of multiple-choice questions involving computation of \(z\), probability, and judgments as to the effectiveness of a training program. These questions review topics and informally introduce hypothesis testing concepts (i.e., judgments of likely and not likely outcomes). Incorrect answers correspond to typical errors, such as failing to consider sample size for standard error calculations or choosing the wrong area under the normal curve. Incorrect answers lead to feedback that addresses why the answer is incorrect and provides guidance for obtaining the correct answer. For example, one question asks the student to choose the probability of obtaining a certain \(z\) score.
Students then draw samples using an interactive applet that graphically represents the population distributions and sampling distributions for the training and no-training groups. Using the applet, the student can manipulate the population mean, sample size, and standard deviation. The student can observe changes in these values immediately in the display of the sampling distributions. However, this exercise asks the student to modify only the means to represent the various training programs. The student begins by drawing a single sample. The applet plots the distribution of the individual scores in the sample and presents the calculated values for the mean and $z$. The student then compares the obtained $z$ to a criterion ($z = \pm 1.96$) and draws a conclusion regarding the null hypothesis. A brief description of decision criteria presents alpha as a standard value for determining whether a sample result is unlikely given that the null hypothesis is true. The student draws 19 more samples, indicating a decision for each sample. For the first exercise, the student examines the highly effective training program, for which most sample means lead to rejection of the null hypothesis.

Next several multiple-choice questions focus on formal aspects of hypothesis testing. One problem provides the mean and $z$ for a sample drawn from one of the training programs. The student must correctly determine the null hypothesis and statistical conclusion. These questions combine with the first sampling exercise to provide the student with a formal understanding of the terminology (e.g., alpha), logic (e.g., low probability suggests null hypothesis is unlikely), and mechanics of hypothesis testing procedures (e.g., reject null hypothesis when $z$ exceeds criterion).

Following these questions are two additional exercises. One exercise examines results for samples taken from a population that differs moderately from the null population (medium effect size). The final exercise uses a population that differs only slightly from the null population (small effect size). Again, the student draws 20 samples, records means, and makes decisions regarding the null hypothesis for each sample.

After completing the sampling exercises, students answer follow-up questions that do not involve computer-based feedback. These questions are appropriate for grading or discussion. One question asks students to examine differences between the frequency of hypothesis rejection for the small and large effect size examples. The student comments on differences, indicating that the example with the larger effect size yielded more rejections of the null hypothesis. The problem then asks the student to suggest reasons for these differences. This type of question requires the student to think about the relationships between the two distributions and consider plausible reasons for different rejection rates. Another question asks the student to evaluate a situation in which a group of program graduates requests a refund from one of the training programs. The graduates claim that the average of their scores was so low on the standardized test that the program’s claims of a certain mean score for graduates could not be true. The student must apply hypothesis testing principles to address the probability that a sample of students would obtain a mean value that deviates from the population mean by a specific amount and discuss the implications of this result for the training program.
and normal distribution probabilities. Items came from the test bank of the textbook that provided the standard laboratory questions. The use of two equivalent versions of the quiz minimized the potential for dishonesty. Students in both groups received either one of the two quizzes randomly. The two quizzes yielded comparable scores, $F(1, 23) < 1$.

**Ratings-Only Group**

Students in the ratings-only group ($n = 112$) responded to questions regarding how easy the tutorial was to understand, how clear they found explanations of statistical concepts, how useful they viewed the tutorial, and their desire to use similar assignments in the future.

**Results**

**Effectiveness Assessment: Comparison With a Standard Laboratory Assignment**

An ANCOVA, controlling for student grade percentage (arc sine transformed) prior to laboratory completion, found that the tutorial group (adjusted $M = 7.50, n = 15$) performed better than the standard laboratory group (adjusted $M = 6.14, n = 10$), $F(1, 22) = 3.96, p = .06$ (two-tailed); $\eta^2 = .15$. This analysis allowed us to control for some systematic differences between students enrolled in different laboratory sections, as the first author observed an unequal distribution of the courses’ top students between conditions. A test of the homogeneity of regression assumption revealed similar relations between current grade and performance for the tutorial and standard laboratory groups, indicating that ANCOVA was appropriate for these data. Although this result did not reach traditional levels of statistical significance, the effect size was encouraging and the two-tailed approach was conservative. Student reports of time spent completing the assignment indicated no substantial difference between the tutorial group ($M = 117$ min) and the standard laboratory ($M = 107$), $F(1, 22) < 1$.

All students using the tutorial judged it as easy to use, compared to only 30% of the students who judged the standard laboratory assignments, Fisher’s exact test, $p < .001$. Students who used the hypothesis testing tutorial also indicated greater interest in using similar assignments in the future (71% very interested, 29% somewhat interested) than students who completed the standard laboratory (30% very interested, 60% somewhat interested, 10% not interested), $\gamma = -0.72, p = .02$.

**Ratings-Only Group**

The second group of students rated the tutorial as somewhat or very easy to use (83%), the explanation of statistical concepts as somewhat or very clear (86%), judged the tutorial as somewhat or very useful for teaching statistics (95%), and were somewhat or very interested in using similar assignments to learn about other statistical topics (94%). Intermediate statistics students rated the tutorial as more useful, $\gamma = 0.72, p < .001$, and indicated greater interest in using tutorials to learn additional topics, $\gamma = -0.71, p < .001$, than did introductory statistics students.

**Discussion**

A test of comprehension suggested that our Web-based tutorial may be more effective than a standard laboratory assignment in teaching basic concepts of NHST. Students rated the tutorial as easier to use and expressed more interest in using similar assignments than students who completed the standard laboratory. Additionally, most students who used the tutorial viewed the explanation of statistical concepts as clear and useful. This combination of increased learning, student interest, and ease of use supports the effectiveness of the tutorial.

The hypothesis testing tutorial incorporates techniques for enhancing learning that may account for improved performances. Bjork (1994) reported that learners often believe that they understand concepts better than they really do. Multiple-choice questions with options that correspond to common mistakes can promote confrontation of misconceptions. When students make mistakes, the tutorial provides instruction that immediately addresses specific misunderstandings. Another potential problem in instruction is a failure to engage students in elaborative processing (i.e., thinking about topics; Hofer, Yu, & Pintrich, 1998). Lack of elaborative processing may be problematic in statistics courses if instruction focuses exclusively on process. Our follow-up questions promote elaborative processing by asking students to explain and apply the concepts they have learned. For example, one question asked students to address the complaints of people who have completed a training course but performed poorly on a standardized test. To address this complaint the student must apply hypothesis testing concepts. Our tutorial may also enhance learning through multimedia presentation. Information presented in multiple formats improves memory (Paivio, 1971). Our interactive applet presents information using text by giving the mean and $\bar{z}$ and graphically by plotting the mean of each sample in relation to the null and true distributions.

Several limitations temper these conclusions. Our assessment does not establish whether use of the computer, opportunity for feedback, or interactive content led to improved performance. Additionally, the sample for the assessment is small. Finally, it is unclear whether students only learn more initially or if our tutorial leads to long-term learning.

However, our hypothesis testing tutorial demonstrates that Web-based materials can incorporate important principles of good instruction. Given the encouraging results of this assessment, we suggest that interactive computer-based tutorials may effectively supplement traditional assignments.

**References**


Notes

1. Grants from the Mellon Foundation, Claremont Graduate University, and Humboldt State University support the Web Interface for Statistics Education.
2. Send correspondence to Christopher L. Aberson, Department of Psychology, Humboldt State University, Arcata, CA 95521; e-mail: CLA18@humboldt.edu.
Using Monte Carlo Software to Teach Abstract Statistical Concepts: A Case Study
Holly Raffle and Gordon P. Brooks
Teaching of Psychology 2005 32: 193
DOI: 10.1207/s15328023top3203_12

The online version of this article can be found at:
http://top.sagepub.com/content/32/3/193

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Society for the Teaching of Psychology

Additional services and information for Teaching of Psychology can be found at:

   Email Alerts: http://top.sagepub.com/cgi/alerts
   Subscriptions: http://top.sagepub.com/subscriptions
   Reprints: http://www.sagepub.com/journalsReprints.nav
   Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Jul 1, 2005
What is This?
Violations of assumptions, inflated Type I error rates, and robustness are important concepts for students to learn in an introductory statistics course. However, these abstract ideas can be difficult for students to understand. Monte Carlo simulation methods can provide a concrete way for students to learn abstract statistical concepts. This article describes the MC4G computer software (Brooks, 2004) and the accompanying instructor's manual (Raffle, 2004). It also provides a case study that includes both assessment and course evaluation data supporting the effectiveness of Monte Carlo simulation exercises in a graduate-level statistics course.

The learning activity described in this article uses the computer program named MC4G: Monte Carlo Analyses for up to 4 Groups (Brooks, 2004). MC4G operates within the Microsoft Windows 9x/NT/XP operating systems. The program performs Monte Carlo robustness, power, and sample size analyses for two, three, and four groups. For two groups, the program performs the overall ANOVA $F$ test, pooled variance independent groups $t$ test, separate variance independent groups $t$ test, and Levene’s test of the homogeneity of variance assumption. For three groups, there are four analysis options: (a) omnibus ANOVA, Welch, and Brown–Forsythe $F$ tests with pooled $t$ test contrasts and adjusted alpha $t$ test contrasts; (b) pairwise and orthogonal $t$ test contrasts; (c) all possible pairwise $t$ test contrasts; and (d) Tukey–Kramer and Scheffé contrasts. For the four-group option, the program calculates the omnibus ANOVA and provides results for several types of contrasts. For all analyses, options exist to change the alpha level, conventional effect sizes, and certain characteristics for data generation. All $t$ tests can be calculated using either mean squared error or separate standard error estimates for each contrast.

Users can save on-screen summary results of the Monte Carlo simulations (see Figure 1) to a disk file for printing from any text editor or word processor. These results include the mean and the standard error for each group, average contrast values, number of rejections, and proportions of rejections from the total number of simulations. The program allows users to save data from individual simulations for later use with MC4G or other statistical packages.

Figure 1. Main MC4G program screen with results displayed from a robustness analysis.
About the Instructor’s Manual

The MC4G instructor’s manual (Raffle, 2004) contains six learning activities. The learning activities cover several beginning and intermediate topics including sampling error, sampling distributions, Type I error, the robustness of statistical tests, and sample size determination. The activities included in the manual use guided discovery learning to encourage students to explore statistical concepts and to draw their own conclusions based on results from MC4G. Ornord (2003) noted that students gain greater conceptual understanding of subject matter when they explore concepts firsthand than when they learn through traditional expository methods. The activities included in the instructor’s manual, like many other activities that use computer simulation methods to teach statistics, are grounded in constructivist theory, which emphasizes the active role of the learner (Mills, 2002; Woolfolk, 2001). The combination of MC4G and the learning activities promotes higher level thinking skills and provides concrete experiences to help students conceptualize abstract statistical ideas.

The instructor’s manual (Raffle, 2004) consists of activities suitable for in-class demonstrations as well as handouts that allow students to work independently, or in small groups, with a personal computer. In addition to the activities in the instructor’s manual, instructors can develop their own lessons on MC4G. For example, after completing the robustness exercises based on the results for the F statistic, students can perform the same exercises using the MC4G screen that provides results for the Welch F and the Brown–Forsythe F statistics and then compare the results. Instructors can also develop exercises for other types of Type I error investigations (e.g., ANOVA vs. multiple t tests) as well as for power explorations.

Case Study

Data

This study used existing data resulting from normal classroom practice. The Ohio University Institutional Review Board determined that this project was exempt from review. We examined two types of data from 23 students enrolled in a graduate-level intermediate statistics course: assessment data and course evaluation data.

Materials

The assessment data consisted of a teacher-made quiz and a teacher-made exam, each containing, among other items, five items on the topic of the robustness of the ANOVA F statistic and Type I error. Students completed the exam approximately 3 weeks after the quiz. The quiz included three true–false items and two multiple choice items relevant to the topic; the exam included five relevant multiple choice items, each with four response alternatives. The instructor included both assessments in the calculation of the students’ course grades.

The course evaluation data consisted of a short questionnaire, including four yes–no response questions and four open-ended questions, in which students evaluated the MC4G program and the corresponding learning activity. The primary, expressed purpose of the questionnaire was to identify ways for the instructor to improve the computer program and the learning activity.

Procedure

After reading textbook content and hearing a classroom lecture about the ANOVA assumption of homoscedasticity, students taking a graduate-level statistics course completed a scheduled quiz. After the quiz, the instructor encouraged students to complete the self-paced, guided discovery learning activity. The MC4G learning activity allowed students to explore the robustness of the ANOVA F statistic, specifically the assumption of homoscedasticity. After completing the learning activity, 21 of the 23 students enrolled in the course completed the short course evaluation form for extra credit, which they submitted to the instructor prior to the exam. The instructor did not provide any additional review of the robustness of the ANOVA F statistic and Type I error topics in class between the quiz and the exam.

Results

Prior to data analysis, we examined the difference scores for accuracy of data entry, missing values, and whether the assumptions of the statistical test were tenable. We conducted a paired-samples t test to evaluate whether students’ knowledge of Type I error and the robustness of ANOVA F statistic increased after completing the self-paced guided discovery learning activity using MC4G. The results indicated that the mean exam score (M = 3.52, SD = 1.24) was significantly greater than the mean quiz score (M = 2.61, SD = 1.31), t(22) = 3.43, p = .002. The standardized effect size index, d, was 0.71. The 95% confidence interval for the mean difference between the two scores was 0.37 to 1.46.

To assess students’ attitudes toward the use and effectiveness of computer simulation activities in the classroom, we used descriptive statistics to analyze the responses to the yes–no items on the evaluation form. Student responses showed overwhelming support for the MC4G learning activity. All 21 students responded “yes” to the first question, “Do you believe that using the learning exercises with MC4G helped you learn the consequences of violating the ANOVA assumption of homogeneity of variances better than what you learned in class alone?” Similarly, all 21 students indicated that they were able to “follow the exercise pretty well.” Twenty students answered affirmatively to the third item, “Would you recommend that I [the instructor] use this exercise in class in the future?” Finally, 20 students answered “yes” to the question concerning whether students should complete a similar exercise for the t test in the prerequisite introductory statistics course.

To analyze the students’ responses to the open-ended items, we used content analysis to search for recurring words, phrases, concepts, and themes (Patton, 2002). When asked...
whether the MC4G learning activity helped to solidify the concept of Type I error and the robustness of the ANOVA F statistic, several students mentioned that the learning activity helped to “make abstract ideas concrete” and to give the learner a “hands-on” opportunity to experience the theoretical material. Students also commented positively regarding the clarification “in practical terms of what is meant when F is described as a liberal or conservative test.”

Students also gave feedback regarding difficulties they encountered with the MC4G learning activity. Although most students noted that they did not have any difficulty with either MC4G or the learning activity, 3 students commented that the learning activity was a helpful adjunct to MC4G. These students communicated that “it would take a while for me to figure out the logic of the program without instruction” and that “the interface of the program is quite intimidating at first glance.” Students also noted that they used class notes and the course textbook while completing the exercise to “provide clarification.”

Students made several general comments and suggestions about the instructor’s future use of the MC4G learning activity. For example, students recommended the inclusion of a conclusion or summary section in the learning activity to help students synthesize the material. Although some students noted that the repetitive nature of the learning activity was helpful, more students advised that the instructor should make the sections of the learning activity less repetitive. Finally, concerning MC4G itself, students recommended that the program should include a graphical explanation of the results and should add a help menu to provide more in-depth explanations.

Conclusions

The purpose of this case study was to determine the effectiveness of Monte Carlo simulation exercises in a graduate-level statistics course. The results of the study indicate that MC4G and the learning activity associated with it were effective in classroom practice. The assessment data indicated that students’ scores after using MC4G and completing the learning activity were significantly higher than students’ scores before the intervention. This finding suggests that the MC4G learning activity was effective in increasing students’ knowledge about Type I error and the robustness of the ANOVA F statistic. Indeed, the course evaluation data point to the fact that MC4G and the learning activity were key reasons why students’ knowledge increased. Student evaluations of both the program and learning activity were overwhelmingly positive. Students indicated that MC4G and the learning activity helped them to understand concretely the abstract theoretical concepts surrounding Type I error rates and robustness. Based on student recommendations, however, instructors should consider providing at least a brief introduction to the computer program in class before having students use the program with discovery learning activities.

References


Notes

1. MC4G, the learning activity, and the accompanying instructor’s manual are available for download free of charge through the author’s Web site: http://oak.cats.ohiou.edu/~brooksg/mc4g.htm
2. Send correspondence to Gordon P. Brooks, Department of Educational Studies, Ohio University, 305A McCracken Hall, Athens, OH 45701; e-mail: brooksg@ohio.edu.
The Utility of Computer-Assisted Power Analysis Lab Instruction
John V. Petrocelli
*Teaching of Psychology* 2007 34: 248
DOI: 10.1080/00986280701700326

The online version of this article can be found at:
http://top.sagepub.com/content/34/4/248
The Utility of Computer-Assisted Power Analysis Lab Instruction

John V. Petrocelli
Wake Forest University

Undergraduate students (N = 47), enrolled in 2 separate psychology research methods classes, evaluated a power analysis lab demonstration and homework assignment. Students attended 1 of 2 lectures that included a basic introduction to power analysis and sample size analysis. One lecture included a demonstration of how to use a computer-based power analysis calculator, whereas the other lecture did not. Students then completed a homework assignment (with or without the power analysis calculator). Compared to students who did not use the power analysis calculator, students who did reported that the lab and homework assignment increased their understanding of power analysis and increased their overall level of interest in their research projects. These students also recommended using the lab in future courses.

Students conducting research, or enrolled in research design and methods courses, need to know how much data they need to collect for their projects to demonstrate a significant effect, if there is one. A power analysis is useful for this purpose, but power analysis calculations can be time-consuming when using complex designs. However, computer-based power analysis calculators that can simplify the task of determining an adequate sample size are readily available. My purpose was to determine whether a brief introduction to the basics of power analysis with the aid of a computer-based power analysis calculator, compared to basic power analysis instruction on its own, enhanced students’ interest in their research projects and increased their intentions to conduct power analyses for future research projects.

In light of research that suggests that use of computer-assisted instruction (CAI) can enhance the learning experience (Traynor, 2003; Worthington, Welsh, Archer, Mindes, & Forsyth, 1996), I hypothesized that basic instruction on power analysis would have a greater impact on student interest and understanding when paired with a power analysis calculator demonstration. My thinking was also consistent with suggestions to “modernize” the classroom with CAI and to prompt students to become more active with new technology (Erwin & Rieppi, 1999; Ralston & Beins, 1999).

Method

Participants

A total of 47 students from two psychology research methods classes attended one of two separate lectures on power analysis (either with a power analysis calculator demonstration or without it) and completed a brief homework assignment during the midpoint of an academic semester. I presented the power analysis lecture to each class 1 week prior to the homework assignment due date.

Power Analysis Calculator

GPOWER (Faul & Erdfelder, 1992) is a computer-based calculator that computes power analyses and sample-size analyses for a wide range of statistical tests. The output of a GPOWER power analysis provides users with feedback on their input as well as information that is useful for understanding the relationships among various concepts essential to power analysis (e.g., sample size, significance level, power, effect size).

GPOWER is an independent program that does not require the user to install any other statistical software.
As a useful teaching tool, GPOWER graphs the relation between any two of the relevant power analysis components. For instance, the relation between effect size and power can be plotted for any statistical test. Students can learn the basic relation between the input components by changing the input and paying particular attention to how minimum sample size changes as a result of changes in effect size, alpha level, or power value and by graphing these components. GPOWER also displays the conventional small, medium, and large effect size values that correspond to the particular test employed (see Figure 1). Finally, GPOWER is free, and students can download it from the Internet (see http://wwwpsycho.uni-duesseldorf.de/aap/projects/gpower/index.html). See Buchner, Erdfelder, and Faul (1997) and Erdfelder, Faul, and Buchner (1996) for additional information regarding technical details and potential uses of GPOWER. There are other power analysis calculators available on the Internet, such as SamplePower 2.0 (SPSS, Inc., 2001), UnifyPow (O’Brien, 1998) and nQuery (Statistical Solutions, 2000). SamplePower 2.0 is an SPSS product; it might be one of the most versatile power analysis calculators available, but with the extensive number of options available, it might confuse novice users. Thus, the program might not be user friendly enough to use in an undergraduate lab demonstration. In addition, SamplePower 2.0 requires a math coprocessor (e.g., SPSS), and the cost is $500. An alternative is UnifyPow, a SAS macro that can be used in conjunction with SAS commands. This power analysis calculator requires the user to own SAS, as well as have some familiarity with SAS programming language. UnifyPow’s “counterpart,” GraphPow, also graphs power analysis results, but it also requires familiarity with SAS programming language and is probably well beyond the scope of what a basic power analysis demonstration requires. Another power analysis calculator, nQuery, is perhaps just as user friendly as GPOWER. With an intuitive interface, it is easy to use and will compute power analyses for all of the designs that GPOWER does, and then some. nQuery also includes the graphing of power components. A downside to nQuery is the cost (over $700). Thus, GPOWER appeared to be the best option with regard to its analysis options, availability, cost, and potential user friendliness.

Procedure

Power analysis instruction. Conducting a power analysis involves some understanding of significance level, effect size, and power. Thus, before proceeding to the description of the assignment, I introduced students to the necessary components in both classes. I introduced students in both classes to basic power sample tables (see Cohen, 1988) and demonstrated how to use the tables to determine the number of participants they would need for seven different scenarios (e.g., “With a significance level of .05, three
groups to compare, a medium effect size of .25, and power of .60, how many participants would be needed to demonstrate a statistically significant effect?"). For each scenario, I first asked students to guess how many participants would be needed and then asked them to use the power sample tables to determine the correct answer.

I introduced students in the lecture with the power analysis calculator demonstration to GPOWER. During the GPOWER demonstration, I highlighted the operations of the calculator (including how to change the test, how to calculate effect size from a previous article using the GPOWER calculator, and how to set the components needed to calculate the power analysis). These students computed the final three of seven scenarios with the assistance of GPOWER.

Assignment implementation. I instructed students to complete a homework assignment worksheet by the following week. The worksheet consisted of five questions. The first three questions varied in the input components of the power analysis, similar to the in-class examples. Students then answered additional questions related to their projects (e.g., how many independent variables their project included, how many levels each independent variable included, what they planned to use as their dependent variable). I required students to calculate the effect size of a statistical test reported in a previous article that was relevant to their research project. I also instructed students to use that effect size in conducting a power analysis for their project, as well as to conduct three additional power analyses with small, medium, and large effect sizes (holding the significance level at .05 and power at .80). I provided students in the class that received the GPOWER demonstration with instructions on how to download the calculator from the Internet. During the following week, I discussed the homework assignment with each class. Students then evaluated the power analysis lab. Finally, I introduced GPOWER to the students in the class that had not received the demonstration the week prior. All students received the same amount of credit for the homework assignment as it was graded as pass or fail.

Results

I conducted a power and sample size analysis for this study. With a large effect size ($d = .80$), alpha level = .05, and power = .80, the total number of participants needed to demonstrate a statistically significant effect in the current study was 42. However, because the actual number of participants totaled 47, the actual power of each $t$ test was .85.

On average, students from the class without the GPOWER demonstration ($M = 2.08$ min) and students from the class with the GPOWER demonstration ($M = 2.33$ min) did not differ in their reports regarding the amount of time it took to complete

Table 1. Student Evaluations of the Power Analysis Lab and Homework Assignment

<table>
<thead>
<tr>
<th>Evaluative Item</th>
<th>With GPOWER</th>
<th></th>
<th>Without GPOWER</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M$</td>
<td>$SD$</td>
<td>$M$</td>
<td>$SD$</td>
</tr>
<tr>
<td>The Power Analysis Lab was appropriate for this course.</td>
<td>6.25</td>
<td>0.94</td>
<td>5.73</td>
<td>1.28</td>
</tr>
<tr>
<td>I now understand how to determine the number of participants I will need</td>
<td>5.87</td>
<td>1.30</td>
<td>4.47</td>
<td>2.15</td>
</tr>
<tr>
<td>for my research project.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Power Analysis Lab increased my interest in my project.</td>
<td>4.40</td>
<td>1.38</td>
<td>3.43</td>
<td>1.56</td>
</tr>
<tr>
<td>I would recommend using the Power Analysis Lab in future sections of this</td>
<td>5.75</td>
<td>1.29</td>
<td>4.86</td>
<td>1.79</td>
</tr>
<tr>
<td>course.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I would run a power analysis for my future research projects.</td>
<td>5.58</td>
<td>1.28</td>
<td>4.78</td>
<td>1.93</td>
</tr>
<tr>
<td>The GPOWER calculator was user-friendly and easy to learn how to use.</td>
<td>6.54</td>
<td>0.72</td>
<td>5.41</td>
<td>1.83</td>
</tr>
<tr>
<td>The GPOWER calculator helped me to learn how to use power analysis to</td>
<td>6.33</td>
<td>0.96</td>
<td>5.41</td>
<td>1.83</td>
</tr>
<tr>
<td>determine the number of participants I would need for an experiment.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I think the GPOWER calculator is a useful way to learn about power analysis.</td>
<td>6.12</td>
<td>0.99</td>
<td>5.41</td>
<td>1.83</td>
</tr>
<tr>
<td>I would recommend using the GPOWER calculator to determine the number of</td>
<td>6.54</td>
<td>0.78</td>
<td>5.41</td>
<td>1.83</td>
</tr>
<tr>
<td>participants needed for an experiment.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Students rated items on a scale ranging from 1 (strongly disagree) to 7 (strongly agree).

$a_n = 24$, $b_n = 23$.

*p < .05.
the power analysis section of the homework assignment, t(45) = 1.35, ns. The two classes did not appear to differ in their ability, as there was no difference between their average, overall grades for the course, t(45) = .19, ns. Evaluations of the power analysis lab and GPOWER appear in Table 1. In general, students of both classes reported that the power analysis lab was appropriate for the course. However, GPOWER students reported that they had a greater understanding for how to determine the number of participants needed for their research projects and a stronger endorsement for using the power analysis lab in future sections of the course than did students who did not receive the introduction to the GPOWER calculator. GPOWER students also reported that the power analysis lab increased their interest in their research projects more than non-GPOWER students. GPOWER students failed to report a significantly greater likelihood of conducting a power analysis for their future projects than the non-GPOWER students; however, the means were in the direction expected.

Students reported that the GPOWER calculator was user friendly and that it was helpful in determining the number of participants they might need for their project. These students also reported that the calculator was a useful way to learn about power analysis and that they would recommend using the calculator for determining how many participants they would need when conducting an experiment.

Discussion

Each specific research design is associated with a unique set of equations to determine the necessary sample size. It is unlikely that research methods instructors have the time necessary to teach students power analysis procedures for each of the commonly employed research designs. Using a power analysis calculator can make the process of teaching power analysis easier by reducing some of the technical detail and still providing students with a solution for a wide range of power analysis inquiries.

This investigation highlighted the usefulness of a computer-based power analysis calculator and a homework assignment. In general, students appreciated the utility of the GPOWER homework assignment and endorsed the power analysis lab and assignment and its use in future sections of research methods courses.

GPOWER is very easy to use, but it is important that faculty do not replace standard instruction with “computer-substituted instruction.” A general understanding of the basic components to power analysis is a prerequisite to understanding GPOWER’s full potential. For instructors who use GPOWER to assist in the instruction of power analysis, it is essential to first provide students with clear definitions of the essential components of power analysis.

A power analysis is an important step in the process of research. Wilcox (1998) suggested that the number of interesting findings lost to insufficient use of statistics is alarming. Perhaps this concern can be reduced by adopting efficient ways to reintroduce power analysis into the undergraduate psychology curriculum.

References

Notes

1. I thank Lina Towett, Elise Percy Hall, and Gordon Kato for their helpful insights and comments on an earlier manuscript and Hope Cantrell for her help with the data collection.

2. Send correspondence to John V. Petrocelli, Wake Forest University, Department of Psychology, Winston-Salem, NC 27109; e-mail: petrocjv@wfu.edu.
Screencast Tutorials Enhance Student Learning of Statistics
Steven A. Lloyd and Chuck L. Robertson
Teaching of Psychology 2012 39: 67
DOI: 10.1177/0098628311430640

The online version of this article can be found at:
http://top.sagepub.com/content/39/1/67

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Society for the Teaching of Psychology

Additional services and information for Teaching of Psychology can be found at:

Email Alerts: http://top.sagepub.com/cgi/alerts
Subscriptions: http://top.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Dec 28, 2011
What is This?
Screencast Tutorials Enhance Student Learning of Statistics

Steven A. Lloyd¹ and Chuck L. Robertson¹

Abstract
Although the use of computer-assisted instruction has rapidly increased, there is little empirical research evaluating these technologies, specifically within the context of teaching statistics. The authors assessed the effect of screencast tutorials on learning outcomes, including statistical knowledge, application, and interpretation. Students from four sections of a psychology course in statistics were randomly assigned to a control text tutorial or an experimental video tutorial group and were tasked with completing a novel statistics problem. Previous math experience, math and computer anxiety, and course grades were also controlled. The results demonstrate that screencast tutorials are an effective and efficient tool for enhancing student learning, especially for higher order conceptual statistical knowledge compared to traditional instructional techniques.

Keywords
statistics, screencasting, vodcasting, podcasting

According to the U.S. Census Bureau (2009), 73.5% of people 3 or older live in a household with Internet access, with a rising trend in the use of mobile media devices (61%) and podcast downloading (27%) for people 18–29 (Madden & Jones, 2008), and more than 80% of college students in the United States own at least one portable audio system capable of downloading audio and sometimes video files (Lum, 2006). In addition, the use of classroom technology, including podcasting, vodcasting, and screencasting, is on the rise in higher education. Some institutions have wholeheartedly embraced this technology and have launched massive campaigns to incorporate podcasting into the curriculum with demonstrated success (Fernandez, Simo, & Sallan, 2009).

Podcasting describes a form of downloadable audio files compatible with MP3 players that has been used for many years by institutes of higher learning to deliver or rebroadcast course content and/or supplemental materials (Donnelly & Berge, 2006; Hammersley, 2004). It is associated with numerous positive learning outcomes affecting a wide range of learners across a number of educational settings (i.e., enhanced learning, increased satisfaction, motivation and engagement, and positive impacts on course-related attitudes and anxiety reduction; Evans, 2008; Hew, 2009; McKinney, Dyck, & Luber, 2009).

There is a large gap between learning theory and teaching practices, which is especially evident when the research involves technological innovations directed toward college students (Fernandez et al., 2009). This gap is evident from the paucity of empirical research evaluating the impact of technologies on learning (Fernandez et al., 2009; Hew, 2009), specifically within the context of teaching statistics (Garfield & Ben-Zvi, 2007).

1 North Georgia College & State University, Dahlonega, GA, USA

Corresponding Author:
Steven Lloyd, Department of Psychology, North Georgia College & State University, Dahlonega, GA 30597
Email: salloyd@northgeorgia.edu
The GAISE Report and the National Science Foundation provide suggestions for teachers of statistics geared toward specific learning goals, including promoting statistical literacy and thinking, using real data, promoting active learning, using technology for developing concepts and analyzing data, and using varied assessment (Franklin & Garfield, 2006; Hall & Rowell, 2008). Computer-assisted instruction and web-based technologies can promote active, task-based learning as well as student independence and conceptual learning in statistics, especially when they are grounded in cognitive learning theory (Lovett & Greenhouse, 2000; U.S. Department of Education [DOE], 2009).

Screencasting, defined as capturing what you do on the computer screen with synched audio commentary (Udell, 2004), is a real-time format that can be disseminated as enhanced podcasts or vodcasts and provides a medium for demonstrating algorithms for problem solving, software instructions, and errors while also providing interpretation-based conceptual understanding in an active learning format. Screencasting encourages meaningful learning according to the cognitive theory of multimedia learning, which suggests that multimodal information presented as combinations of narration and animation, when appropriately temporally and spatially sequenced, self-paced, coherently communicated, and stated in a conversational manner, leads to problem-solving transfer in novel situations and encourages active cognitive processing and cognitive load reduction to promote deeper learning (Mayer, Fennell, Farmer, & Campbell, 2004; Mayer & Moreno, 2003).

Given the need for empirical research on new instructional technologies, specifically within the context of statistical instruction, we set forth to determine the impact of a supplemental vodcast tutorial, which was designed according to multimedia learning theory recommendations, on objective learning outcomes in reference to statistical knowledge, application, and interpretation. Our results reveal that vodcasting is an effective and efficient tool for enhancing student learning, especially for higher-order conceptual statistical knowledge.

**Method**

**Participants**

A total of 53 students from four sections of an upper level psychology course in statistics participated. Stratified randomization was used to assign participants to experimental conditions based on gender. The sample consisted of predominantly young (R = 20–50, M = 23.45, SD = 4.83), Caucasian (91%, n = 48), upper level (junior or senior class status; 100%, n = 53), female (81%, n = 43) psychology majors (81%, n = 43) from a public university in the southeastern United States. All participants had taken a prerequisite course in elementary statistics, but the groups did not differ in the number, type, or level of additional math courses they had taken or completed (p > .05).

All procedures were performed in accordance with the university’s institutional review board guidelines.

**Materials and Procedures**

**Psychological scales.** A 10-item Math Anxiety Scale (MAS; Betz, 1978) and a 6-item Computer Anxiety Scale (CAS; Lester, Yang, & James, 2005) were administered during the first week of the semester.

**Screencast tutorials.** A screencast tutorial was created using iShowU (www.shinywhitebox.com) and iMovieMaker and was served as a vodcast. The tutorial demonstrated the following steps of statistical analysis: data entry, conducting an independent samples t test analysis, and working with output files in SPSS. The completed tutorial was 11.55 min in length. The control group was given a packet of material taken from an SPSS user guide, which covered the same content and included screen shots of the same SPSS environment demonstrated in the screencast tutorial (Kirkpatrick & Feeney, 2006). Participants in both groups were given 12 min to review their tutorials.

**Statistical problem set.** The statistical problem set necessitated an independent samples t test analysis. The raw data were listed, and the participants performed analyses using SPSS (v. 16.0). They were given 25 min to complete the task in Experiment 1. In Experiment 2, the time allotted to complete the task was increased to 55 min, and participants were allowed to review the video or text tutorial at the time of testing.

**Scoring the statistical problem set.** There were 10 possible points for this exercise. A point was awarded for correctly reporting the mean (Group 1), standard deviation (Group 1), mean (Group 2), standard error of the mean (Group 2), standard error of the mean difference score, t obtained, and p value. A point was also awarded for correctly rejecting the null hypothesis, using the correct reporting format, and stating the correct conclusion. Screencapture and mousecapture (iShowU) were used to record participants as they solved the problem. The number of mouse clicks executed and the time to complete the assignment were extracted from these recordings.

**Results**

**Experiment 1**

The experimental group did not differ in self-reported computer anxiety, t(29) = 0.886, p = .383, or math anxiety, t(29) = 0.100, p = .851, as measured by the CAS and the MAS, respectively. Although the MAS scores were positively correlated to CAS scores, r(29) = .385, p = .032, neither the MAS score, r(29) = −.198, p = .287, nor the CAS scores, r(29) = .009, p = .963, were correlated with final course grades.

The screencast tutorial group took less time to complete the statistical problem (M = 15.20, SE = 0.70) than the text tutorial group (M = 18.06, SE = 0.67), t(29) = 2.950, p = .006 (Figure 1). The screencast tutorial group also scored higher on the statistical problem set (M = 7.27, SE = 0.30) than did the text tutorial group (M = 4.5, SE = 0.55), t(29) = 4.347, p < .001 (Figure 2). In addition, the time to task completion
negatively correlated with total score on the problem set, $r(29) = -0.510$, $p = 0.003$.

The screencast tutorial group outscored the text tutorial group on several individual questions, including correctly stating the mean (Group 2; $M = 1.00$, $SE = 0$ and $M = 0.75$, $SE = 0.11$, respectively), $t(29) = 2.163$, $p = 0.039$, the $t$ obtained ($M = 1.00$, $SE = 0$ and $M = 0.38$, $SE = 0.13$, respectively), $t(29) = 4.836$, $p < 0.001$, the $p$ value ($M = 0.73$, $SE = 0.12$ and $M = 0.25$, $SE = 0.11$, respectively), $t(29) = 2.973$, $p = 0.006$, and the standard error of the mean difference score ($M = 1.00$, $SE = 0$ and $M = 0.50$, $SE = 0.13$, respectively), $t(29) = 3.746$, $p = 0.001$, and correctly rejecting the null hypothesis ($M = 0.47$, $SE = 0.13$ and $M = 0.06$, $SE = 0.06$, respectively), $t(29) = 2.802$, $p = 0.009$ (Table 1). The ability to correctly reject the null hypothesis was positively correlated with the total score on the problem set, $r(29) = 0.412$, $p = 0.021$.

Experiment 2

Given additional time to solve the problem (55 min vs. 25 min in Experiment 1) and the ability to review the video or text tutorial at the time of testing, the screencast group ($M = 7.27$, $SE = 0.70$) outperformed the text tutorial group ($M = 5.36$, $SE = 0.85$), $t(20) = 2.15$, $p = 0.044$ (Figure 2), but the time to task completion did not differ between the video group ($M = 38.00$, $SE = 2.92$) and the text group ($M = 34.27$, $SE = 3.33$), $t(20) = 0.841$, $p = 0.41$ (Figure 1). Differences in specific test questions included the ability to correctly state the value of $t$ obtained, $t(20) = 2.39$, $p = 0.027$, and the standard error of the mean difference score, $t(20) = 2.39$, $p = 0.027$ (see Table 2).

The most efficient strategy to solve the statistical problem was predetermined. Neither the video tutorial group ($M = 229.80$, $SE = 26.14$) nor the text tutorial group ($M = 228.67$, $SE = 29.81$) used efficient strategies, nor did they differ from one another, $t(20) = 0.029$, $p = 0.98$.

Discussion

Despite the rise in online instruction, there are few empirical, methodologically sound studies assessing web-based
data suggest that the video tutorial group were not just following algorithms based on rote memorization but that their demonstrated enhanced learning arose from better conceptual understanding and problem-solving transfer.

This study provides support for the use of computer-assisted technology in teaching statistics to undergraduate psychology students. These results are extremely relevant given the challenges that instructors face in teaching statistics, especially considering its importance in the undergraduate psychology curriculum. Future studies should consider whether the use of vodcasting provides the same benefits when used to supplement an entire course, perhaps in out-of-class, nonproctored labs or asynchronous online environments.

Acknowledgments
The authors wish to acknowledge the following people for their assistance in the preparation of this article: Catherine Ashley, Kelly Cate, and Ashley Marascalco.

Declaration of Conflicting Interests
The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The authors disclosed receipt of the following financial support for the research and/or authorship of this article: This research was supported by grants from the NGCSU QEP and the NGCSU CTLE.

References


Table 2. Experiment 2

<table>
<thead>
<tr>
<th>Measure</th>
<th>Screencast</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N M SE</td>
<td>M SE t(20) p</td>
</tr>
<tr>
<td>Time to complete (min)</td>
<td>22 38.00 2.92</td>
<td>34.27 3.33 0.84 &gt;.05</td>
</tr>
<tr>
<td>Total score (out of 10)</td>
<td>22 7.27 0.27</td>
<td>5.36 0.85 2.15 .044</td>
</tr>
<tr>
<td>Specific questions&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct Group 2 mean difference</td>
<td>22 1.00 0</td>
<td>0.91 0.09 1.00 &gt;.05</td>
</tr>
<tr>
<td>Correct SE of the M</td>
<td>22 1.00 0</td>
<td>0.64 0.15 2.39 .027</td>
</tr>
<tr>
<td>Correct t value</td>
<td>22 1.00 0</td>
<td>0.64 0.15 2.39 .027</td>
</tr>
<tr>
<td>Correct p value</td>
<td>22 0.73 0.14</td>
<td>0.36 0.15 1.75 &gt;.05</td>
</tr>
<tr>
<td>Correct rejection of null</td>
<td>22 0.45 0.16</td>
<td>0.09 0.09 2.00 &gt;.05</td>
</tr>
</tbody>
</table>

<sup>a</sup> Correct response = 1, incorrect response = 0.

According to the cognitive theory of multimedia learning, the mind is a dual-channel, limited-capacity, active processing system, which benefits from learning strategies utilizing cognitive load reduction and multimodal learning, especially for semantic memory encoding and accessibility for working memory in problem-solving transfer (Mayer, 2001). Automaticity, increasing what you know about the top, and developing and working through a systematic plan also improve problem solving (Ashcraft, 2006). Empirical evidence supports a multimedia effect for retention and problem-solving transfer.


Improving the Science Education of Psychology Students: Better Teaching of Methodology

Robert W. Proctor and E. J. Capaldi

Teaching of Psychology 2001 28: 173
DOI: 10.1207/S15328023TOP2803_02

The online version of this article can be found at:
http://top.sagepub.com/content/28/3/173

Published by:
SAGE
http://www.sagepublications.com

On behalf of:

Society for the Teaching of Psychology

Additional services and information for Teaching of Psychology can be found at:

Email Alerts: http://top.sagepub.com/cgi/alerts

Subscriptions: http://top.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Jul 1, 2001

What is This?
Our examination of a number of widely used methodology texts in psychology revealed that they present a narrow conception of science that fails to recognize major changes in the understanding of science that have occurred since 1960. Such changes include an emphasis on explanation, theoretical promise and scientific importance, and the difficulties associated with hypothesis testing. Because these texts provide the primary source of science education for psychology students, they poorly prepare students for dealing with contemporary science. In this article, we detail the inadequacies of current methodological texts with respect to their treatment of science and describe a more currently acceptable conception of science. We provide instructors with recommendations for improving the science education of psychology students.

The much sought-after “scientific method” may be a will-o’-the-wisp. To seek the rules of the scientific method is to presuppose that there is only one legitimate means to the achievement of the shared cognitive aims of science.

—Laudan, 1984, p. 36

But is there such a thing as the scientific method? The answer is clearly no … the methods of science are complex, variegated, and often local.

—Oreskes, 1999, p. 5

Textbooks are clearly an important overall component of sound and effective teaching. To the extent that a textbook is deficient in some respect, correction of that inadequacy by some means must occur, most notably by the direct intervention of the instructor using it. Such direct teacher intervention appears to be required with respect to current methodology texts in psychology, notably where the intent of the scientific method is concerned.

Virtually all undergraduate psychology majors must take a course dealing with research methods in psychology. In Perlman and McCann’s (1999) review of 500 college catalogs for four institutional types, 78% of the departments required a course in research methods or experimental psychology for undergraduate majors, second only to the requirement of introductory psychology by 93% of the departments. For many undergraduate students this course is their major, if not sole, experience with research methodology. Although numerous research methods texts are available, our examination of many of the widely used texts indicates that their coverage tends to be largely similar. Typically, they contain one or more chapters describing the scientific method, a chapter concerned with research ethics, a number of chapters devoted to specific methodological issues in conducting behavioral research, and several chapters dealing with statistics. In treating the nuts and bolts of research methodology, for example, experimental design and certain procedural matters such as double-blind experiments, experienced researchers would probably agree that the coverage is informative to neophytes.

There is a major, extremely important respect, however, in which these texts are seriously deficient: They present an outdated, and thus incomplete, depiction of scientific practice that is misleading to students. Because this depiction is antihistorical, it ignores contemporary scholarship revealing that science as actually practiced differs considerably from the outdated and passé conception of science presented in these methodology texts. Thus, although students who have mastered the material in these textbooks will have an adequate idea of procedural matters such as experimental design and control, they will have an inadequate understanding of science and the practice of science in the contemporary world.

The main deficiency of these research methods texts in terms of representing science is that they do not reflect the best thinking about science that has occurred from about 1960 to the present: Not only are they mired in the past, but they treat science itself as a finished, completed project, rather than a work in progress that is becoming increasingly better understood. More specifically, methodology texts overemphasize the role of hypothesis testing in science and underestimate the role of explanation, theoretical promise, and scientific importance. This article reveals the deficiencies of current methodological texts with regard to contemporary conceptions of science and suggests various ways that instructors can provide students with a more adequate understanding of how science is actually practiced. These deficiencies are best understood against the background of three major methodological movements that have occurred in science.

Three Great Methodological Eras in Science

Excluding views that predate the Newtonian era, there have been three great methodological eras in science. The
first, the era of induction, began in the late 17th century and extended until about the middle of the 19th century. In this era, the objective of science was to produce correct inductions by the patient and careful observation of phenomena. That is, the intent was to arrive at correct generalizations by examining specific cases. Francis Bacon and subsequently Isaac Newton advocated this approach. Newton said of his scientific system in a letter to Cotes in 1713, “These principles are deduced from phenomena and made general by induction, which is the highest evidence that a proposition can have in this philosophy” (Thayer, 1953, p. 6). Philosophers and scientists alike accepted the inductive approach for the next 150 years. The great popularizer of Newton’s point of view in the 18th century was Reid (1785/1969), who said,

[Scientific] discoveries have always been made by patient observation, by accurate experiments, or by conclusions drawn by strict reasoning from observations and experiments, and such discoveries have always tended to refute, but not to confirm, the theories and hypotheses which ingenious men have invented. (p. 44)

In this era, scientists favored induction and eschewed hypothesis testing and the postulation of unobservable entities (e.g., see Encyclopaedia Britannica, 1771, p. 34).

In the second great methodological era, extending from about 1850 to 1960, induction was considered to be inadequate for the task of creating adequate scientific theories. In this period, scientists and philosophers suggested two great additions to induction: hypothesis testing and the postulation of unobservable entities such as atoms. The great innovator and popularizer of hypothesis testing as the proper approach to science was Whewell. According to Whewell (1847/1967), “The process of scientific discovery is cautious and rigorous, not by abstaining from hypotheses, but by rigorously comparing hypotheses with facts, and by resolutely rejecting all which the comparison does not confirm” (p. 468). Essentially, Whewell suggested that the hallmark of successful scientific activity is the confirmation of predictions derived from theory or experience. The two great philosophies of science in this period, those of the logical positivists and of Popper, accepted Whewell’s approach. According to logical positivists, the important aspect of science was the confirmation of hypotheses by repeated tests, an inductive approach. However, in Popper’s (1959) view, the important aspect of science was the attempt to falsify hypotheses, a deductive approach. In this era, as in the era of Newton, foundationism provided the basis for justifying methodology. That is, a researcher judged the adequacy of methodological principles primarily on the basis of logic and intuition. The outstanding example here is Popper. His falsification principle, as Popper himself was quick to acknowledge, was just a convention. That is, Popper devised a methodological principle that, in his opinion, is one that scientists ought to follow.

The third great methodological era in science is a consequence of a major contemporary movement in the philosophy of science called naturalism (see, e.g., Callebaut, 1993). Naturalism is the view that all issues, including issues of scientific methodology, should be evaluated empirically. Thus, naturalism recommends that researchers study science in the same way that they study specific sciences such as biology and psychology. The most well-known proponent of a naturalistic methodology among psychologists is Kuhn (1962). Kuhn arrived at a number of conclusions concerning science as a result of a historical analysis. It is not Kuhn’s specific conclusions about science that concern us here, but rather the historical method he used in arriving at these conclusions. The historical (i.e., empirical) implications of Kuhn’s treatment seem to be totally unappreciated by psychologists. Kuhn’s empirical approach to science stands in stark contrast to the foundationalist approach employed in the preceding two eras. Consider some specific examples. According to the foundationalist approach employed by the logical positivists and Popper, a successor theory should explain not only all that a displaced theory explained but new things as well. Kuhn’s historical scholarship indicated that no theory did this; that is, in relation to displaced theories, successor theories possess both advantages and disadvantages. As another example, one might ask whether historical research reveals that the establishment of all accepted scientific theories has occurred on the basis of a particular method, the hypothetico-deductive method. We have more to say about this later.

Since Kuhn’s (1962) original writings, naturalism has become popular within the philosophy of science itself and, like Kuhn’s writings, is underappreciated by psychologists. Naturalism, we suggest, has a number of desirable implications for the understanding of methodology within psychology, as indeed within science generally (e.g., see Capaldi & Proctor, 1999, 2000). Among the best known contemporary naturalists are Campbell, Giere, Laudan, Mayo, Rosenberg, and Shapere. A good general description of naturalism is available in Callebaut (1993). According to Callebaut,

Naturalism as a philosophical movement claims that whatever exists or happens in the world is susceptible to explanation by natural scientific methods; it denies that there is or could be anything which lies in principle beyond the scope of scientific explanation. Although naturalism is firmly rooted in the philosophical tradition (materialism, empiricism), a thoroughly naturalized philosophy of science is only being developed now. (p. xv)

One example of the empirical approach to methodology suggested by naturalism is found in the writings of Mayo (1996), who recently developed a major approach to methodology. She said of her approach:

Methodological rules do not rest on a priori intuitions, nor are they matters to be decided by conventions (e.g., about what counts as science or knowledge or rational). They are empirical claims or hypotheses about how to find out certain things by arguing from experiments . . . . Hence, the account I propose is naturalistic. (p. 19)

By ruling out a priori intuitions and conventions, Mayo directly ruled out Popper’s (1959) approach to methodology, which is based on both. Mayo’s analysis also rules out the approach to methodology adopted by the logical positivists (see Laudan, 1996) because, according to them, methodological statements are neither synthetic nor analytic.

An outstanding example of naturalistic thinking about scientific methodology appears in Laudan (1996), a position that
McGuigan went on to say, “An understanding of the philosophy of science is considered, the tendency is to mention approvingly some older view, often Popper’s falsificationism, that predates naturalism. An exception to this trend is McBurney’s (1997) text, which mentions Laudan’s (1977) treatment of the problem-solving nature of science.

An individual defending the way in which these texts treat science might give as a major justification that the authors are providing a necessarily simplified, but basically correct, view of science to undergraduates. This justification is inadequate. First, these texts contain no notice whatsoever of the profound changes in the philosophy of science that have occurred since 1960. An examination of the other methodological texts reveals that their treatments of science are much like McGuigan’s in that they ignore recent advances in understanding science. When, rarely, the philosophy of science is considered, the tendency is to mention approvingly some older view, often Popper’s falsificationism, that predates naturalism. An exception to this trend is McBurney’s (1997) text, which mentions Laudan’s (1977) treatment of the problem-solving nature of science.

An individual defending the way in which these texts treat science might give as a major justification that the authors are providing a necessarily simplified, but basically correct, view of science to undergraduates. This justification is inadequate. First, these texts contain no notice whatsoever of the profound changes in the philosophy of science that have occurred since 1960. The absence of such notice could conceivably mean that in the authors’ opinions recent advances in the understanding of science are not worthy of extensive treatment. However, this view does not seem to be the case because these texts contain scant recognition of these recent advances, even to criticize them. In short, these texts de-
scribe an understanding of science as it existed in the second major methodological era noted earlier and have not progressed beyond that. As indicated, the most frequently mentioned philosopher of science in these texts is Popper and his falsification principle. Ironically, treatment of Popper fails to mention entirely that his views have come under enormous criticism within the philosophy of science and have been largely rejected or seriously modified (see, e.g., Kuhn, 1970; Lakatos, 1970; Laudan, 1996; Mayo, 1996; Salmon, 1988).

Alternatives to Hypothesis Testing As Revealed by Recent Historical Scholarship

The standard description of the scientific method in psychological research methods texts is highly similar to that suggested by G. K. Gilbert, a prominent geologist, in 1886. We quote from Gilbert for two major reasons. The first is to indicate that the tendency to identify science with hypothesis testing is not limited to psychology but can be found in other areas of science. The second reason is that a major theoretical development in geology, described later, illustrates a deficiency of identifying the scientific method with hypothesis testing alone. This deficiency is one shared by all areas of science, including psychology. According to Gilbert (1886), who strongly influenced subsequent generations of geologists,

It is the province of research to discover the antecedent of phenomena. This is done by the aid of hypothesis. A phenomenon having been observed, or a group of phenomena having been established by empirical [all] classification, the investigator invents an hypothesis in explanation. He then devises and applies a test of the validity of the hypothesis. If it does not stand the test he discards it and invents a new one. If it survives the test, he proceeds at once to devise a second test. And thus he continues. (p. 285)

The first, and perhaps most obvious, problem with this approach is that in the initial stages of developing a theory it is quite easy to subject it to a procedure that has the capacity to disconfirm it. Subjecting novel theories to strong hypothesis testing in their initial stages of development would result in few surviving, it is safe to say, as Lakatos (1970) showed in some detail. Note that Gilbert (1886), like recent psychological methods texts, failed to acknowledge that hypothesis testing is of limited value in the initial stages of developing hypotheses.

Another difficulty with the approach taken in methodological texts and by Gilbert (1886) is that it suggests that one tests a specific hypothesis in isolation. A more formal statement of the violation involved here is known as the Duhem–Quine thesis (for discussions of this thesis, see Chalmers, 1999; Lakatos, 1970, 1976; O'Donohue & Smith, 1992). In brief, the Duhem–Quine thesis suggests that a number of other matters are under test in addition to the hypothesis itself, such as the reliability of the testing equipment, whether the hypothesis has been validly derived from the theory, and a variety of auxiliary assumptions contained within the general theory from which the hypothesis is derived.

An implication of the Duhem–Quine thesis, which is consistent with common sense, is that a failure to confirm a hypothesis may occur for a variety of reasons, and it may be either difficult or impossible to isolate the source of the disconfirmation. It may be the hypothesis directly under test that is at fault. It may be an invalid derivation from the theory that is at fault. It may be that the failure lies with the testing equipment or the theory that underlies use of the equipment. It may be that the hypothesis is correct, but one or another of the auxiliary assumptions of the theory is incorrect. Contrary to this implication of the Duhem–Quine thesis, methodological texts in psychology imply that it is relatively easy to test hypotheses, either accepting or rejecting them. Methodology texts fail to mention the Duhem–Quine thesis directly. In one of the rare cases in which the thesis was considered indirectly, it was misinterpreted. Shaughnessy et al. (2000) succinctly represented as follows the position of Lakatos (1978), who accepted the Duhem–Quine thesis: “A theory is not likely to fall on the basis of a single test (e.g., Lakatos, 1978)” (p. 31). However, Lakatos’s actual position was that a theory can always evade falsification if its author is intent on making it so (see Lakatos, 1970, pp. 93–116). In the last expression of his views, in 1973, just before his death in 1974, Lakatos (Lakatos & Feyerabend, 1999) clearly indicated that whether the test of a hypothesis is regarded as in any way definitive depends on agreement by convention that the statement is true or false (p. 79). He went on to say that scientists often exercise tenacity in the sense of continuing to accept an apparently falsified hypothesis and that such tenacity has proven to be a good thing in the history of science (p. 89).

When data appear to disconfirm a theory, a common procedure in science is not to reject the theory, but rather to add some auxiliary hypothesis that rescues the theory. Auxiliary hypotheses have an undeserved bad reputation in many quarters. Contrary to common opinion, it is not easy to devise an auxiliary hypothesis that simultaneously saves a theory while failing to produce new conceptual or empirical problems (see Laudan, 1977, pp. 114–118).

A further difficulty with the textbook description of hypothesis testing, also found in Gilbert (1886), is that historical scholarship reveals that scientists seldom attempt to test theories with the intention of falsifying them. Rather, what scientists do most of the time is attempt to confirm their theories by making novel predictions from them. Scholarship seems to reveal that scientists are most impressed with theories that make predictions that are both novel and unexpected (e.g., see Chalmers, 1999; Lakatos, 1976).

If it were the case that failure to confirm a hypothesis would be legitimate grounds for rejecting it or the theory in which it is embedded, as psychology texts and Gilbert (1886) suggest, then rejection of all theories known to science would occur because all theories are falsified by at least some observational data. For example, the orbit of the planet Uranus apparently refuted Newton’s theory. However, scientists at the time did not reject Newton’s theory on that basis, but rather posited that another planet as yet unobserved, Neptune, was the cause of the disturbance in Uranus’s orbit. Of course, observation of Neptune, which ultimately occurred, brought Newton’s theory in line with the observational data. As another example, when Newton published his Principia, it was generally known that his theory could not explain the motion of the moon. As Chalmers (1999) indicated, it took almost 50 years to determine that Newton’s inability to explain the motion of the
moon was due to causes other than deficiencies in the theory itself. Lakatos (1976) and Chalmers (1999) cited a variety of other examples of major theories that were unable to explain a variety of observations at the time they were suggested.

Another major difficulty in identifying hypothesis testing as the scientific method is, ironically, that too heavy a reliance on hypothesis testing may actually be counterproductive, as we suggested earlier in connection with geology. This view is amply illustrated by the rejection by American geologists of plate tectonics long after its acceptance by European geologists. Plate tectonics is the explanation of continental drift that is widely accepted in all of geology (e.g., Lutgens & Tarbuck, 1998). Wegener’s (1924) method, in initially proposing plate tectonics, was to show that it was consistent with a considerable number and variety of already known phenomena. American earth scientists rejected this approach to establishing a theory, which we might call the explanatory approach. They considered that the proper route to establishing a theory was via confirming it through hypothesis testing (see Laudan, 1996). As Oreskes (1999), who examined the controversy over plate tectonics in considerable detail, stated in her recent book:

The thesis of this book is that American earth scientists rejected the theory of continental drift not because there was no evidence to support it (there was ample), nor because the scientists who supported it were cranks (they were not), but because the theory, as widely interpreted, violated deeply held methodological beliefs and valued forms of scientific practice. (p. 6)

A brief explanation of how geologists came to hold these methodological beliefs is in order. Chamberlin (1890/1965), another prominent geologist, agreed with Gilbert (1886) that hypothesis testing was the scientific method. Chamberlin rejected the idea that science could be advanced by employing explanatory theories of the type recommended by Wegener (1924). Earth scientists in the United States widely accepted Chamberlin’s view. Thus, when Wegener propounded his view of plate tectonics, American earth scientists widely rejected it. The European earth scientists, who did not value hypothesis testing as much as their American counterparts, consequently accepted Wegener’s views at a much earlier date.

What the Wegener episode in geology illustrates is that different scientists employ different methodological and empirical criteria in accepting or rejecting theories. Methodology texts in psychology strongly imply, even if they do not state directly, that all scientists should use a single criterion in accepting or rejecting theories, namely that of hypothesis testing. If Wegener were a devotee of hypothesis testing in arriving at a theory, serious consideration of plate tectonics as a theory may not have occurred until a much later time. The fact that different scientists employ different criteria in evaluating theories is highly useful in science. In addition to the Wegener incident, other considerations also illustrate this point. Consider two major examples. Scientists who put forward some novel theory long before it is generally accepted by the field as a whole are, justifiably or unjustifiably, employing criteria different from their colleagues. Often, such novel theories turn out to be incorrect, but sometimes they turn out to be both correct and highly useful. Another place in which it is evident that different scientists employ different criteria in the evaluation of different theories is in the rejection of theories. Well-established theories, when rejected, are not rejected simultaneously by all members of the relevant scientific community. Indeed, as Kuhn (1962) pointed out, some members of the scientific community may continue to accept a theory long after the field has abandoned it. The lesson here seems obvious: It is on the whole beneficial to science that different scientists employ different criteria for accepting and rejecting theories.

Other Scientific Methods

If hypothesis testing is the scientific method, then a straightforward prediction follows: There should be no significant conceptual or empirical advance in science that arises from a method other than hypothesis testing. What we show in the following section is that this prediction is totally false. In other words, it is our intention to show that hypothesis testing is not the only useful method employed in science, a point neglected in current methodology texts.

Induction

As we have already seen, most of the significant scientific theories developed in the 16th to 18th centuries, in the view of the authors of those theories, employed simple induction. Induction, of course, has its place in science, particularly in the beginning stages of investigation. Inductive logic, as distinct from induction as a methodological procedure, is mentioned briefly in some of the current methods texts, including Graziano and Raculin (1997), Goodwin (1998), Kantowitz et al. (1997), McGuigan (1997), Nation (1997), Smith and Davis (1997), and Stangor (1998).

Explanation

There are three separate approaches to demonstrating the importance of explanation as an important scientific method. The three are logical arguments in favor of explanation, specific statements by scientists indicating that explanation was a greater factor in the acceptance of theories than prediction from a hypothesis, and finally, demonstrating that significant scientific theories (in addition to plate tectonics) have been accepted on the basis of their explanatory capacity alone.

Let us consider logical arguments first. A variety of major figures have suggested that explanation is as important a factor as prediction in evaluating scientific claims. John Stuart Mill was one of the earliest figures to take this position, specifically directing his arguments against Whewell. More recently, Lakatos (1970) cited Keynes (1921) on this point, indicating,

The probability of a theory given evidence cannot possibly be influenced … by when the evidence was produced: the probability of a theory given evidence can depend only on the theory and the evidence, and not upon whether the evidence was produced before or after the theory. (pp. 123–124)
Brush (1989) suggested that under some circumstances explanation is superior to prediction. Brush began by asking readers to assume that a well-known phenomenon has resisted explanation from a variety of theoretical viewpoints. Devising an explanation for that phenomenon would be a very important event, as Brush pointed out. More recently, Snyder (1994) made much the same point as Keynes, indicating essentially that the time at which evidence is known relative to the invention of a theory is irrelevant to the evaluation of the theory. Achinstein (1994) argued that under some circumstances, explanation may be better than prediction, as suggested by Brush, but that under other circumstances, prediction may be better than explanation.

Statements by working scientists reviewed by Brush (1989) indicate that they sometimes consider explanation more valuable than prediction. Brush examined most of the published comments by physicists during the first few years after Einstein’s prediction from relativity theory concerning the bending of light was confirmed. He reported that this prediction was not more valued than relativity’s explanation of Mercury’s orbit. In short, he found that almost all published comments, in evaluating Einstein’s theory of relativity, valued explanation at least as much as prediction.

Finally, examples of explanatory theories that have proved highly successful appear in Donovan, Laudan, and Laudan (1992). We mention several specific examples in some detail later in the article.

**Promise**

The recognition of overarching theoretical frameworks called paradigms (Kuhn, 1962), research programs (Lakatos, 1970), or research traditions (Laudan, 1977) is one of the great advances in contemporary philosophy of science. Research traditions contain very general ideas about what is important in the creation of scientific theory. Psychology has embraced a number of research traditions, to cite but a few examples, associationism, behaviorism, and cognitive psychology. Research traditions provide the basis for the development of specific theories, subsequently tested employing various methodological procedures. For example, behaviorism has given rise to a variety of learning theories that claim to be behavioristic, such as Skinner’s (1938) radical behaviorism, Hull’s (1943) methodological behaviorism, and Tolman’s (1932) cognitive behaviorism, among others. Research traditions themselves, when initially advanced, cannot claim in any sense to be firmly established. For example, Watson (1913), in his famous paper advancing behaviorism over structuralism, said:

Certainly the position I advocate is weak enough at present and can be attacked from many standpoints. Yet when all of this is admitted, I still feel that the considerations which I have urged should have a wide influence upon the type of psychology which is to be developed in the future. (p. 175)

The inadequate evidentiary basis for the initial acceptance of behaviorism is not unique; it is characteristic of all general research traditions in the initial stages. How, then, do research programs come to be accepted? The answer appears to be straightforward: They are accepted on the basis of their promise to provide adequate answers for a variety of theoretical issues. That is, when introduced, they seem more promising than their alternatives as the basis for developing adequate scientific theories. Thus, behaviorism, when introduced, seemed more promising to its adherents than did structuralism. By the same token, when the cognitive approach was recommended in the 1960s, it seemed more promising to its adherents than the behavioristic approach that it attempted to displace. Arguably, in terms of introducing viable scientific theories, promise plays as big a role in science as any other single factor. To omit the role of promise from descriptions of science is to omit a spectacularly important feature of science.

**Importance**

A scientist’s decision to work with a particular theory on the basis of its importance is abundantly demonstrated on the contemporary scene by the considerable number of outstanding physicists who are willing to work with string theory (Greene, 1999). Currently, string theory, for a variety of reasons, is unable to deliver a single confirmable prediction. Physicists use string theory because, at present, it is an approach that promises to reconcile quantum mechanics and relativity theory. Here is a case in which both prediction and explanation are foregone because of the importance of the scientific problem. Nothing we have said means that theories in the final analysis should fail to make confirmable predictions. The operable words here are *in the final analysis*. On the basis of historical analysis, the empirical lesson seems to be that theories that have no empirical consequences when introduced may be accepted because of their importance. As Greene (1999), a prominent string theorist, said:

The history of physics is filled with ideas that when first presented seemed completely untestable but, through various unforeseen developments, were ultimately brought within the realm of experimental verifiability. The notion that matter is made of atoms, Paul’s hypothesis that there are ghostly neutrino particles, and the possibility that the heavens are dotted with neutron stars and black holes are three prominent ideas of precisely this sort—ideas that we now embrace fully but that, at their inception, seemed more like the musings of science fiction than aspects of science fact. (p. 226)

**Suggestions for Improving Methodological Instruction**

In our opinion, the first edition of McGuigan’s (1960) book represented a significant watershed in methodology texts. We agree strongly with his assertion that treatments of methodology and of science should pay close attention to developments in the philosophy of science itself. Clearly, the current methodology texts in psychology, including revisions of McGuigan’s original text, have failed to do this. This failure is a serious defect in these texts that deprives students of an adequate appreciation of the best available contemporary science. For teachers to impart a proper appreciation of re-
search methodology to students, they must embed the research design issues within a more accurate, more complex, and nuanced depiction of science. If they fail in this, their students will come away from methodology courses with little more than a cookbook appreciation of how contemporary science is conducted.

Faithfulness with the spirit of science demands that instructors deal with a variety of contemporary issues that the texts currently avoid. Some particularly important examples of issues with which the instructors must deal include the following:

1. In the spirit of the recommendation expressed by McGuigan (1960) in his initial influential text, we suggest that perhaps the most important change that instructors can initiate in contemporary methodology courses is increased coverage of, and emphasis on, the philosophy of science. We mentioned previously many of the most important philosophers of science on the contemporary scene, and instructors may consult them. In addition, we recommend two up-to-date philosophy of science texts that provide excellent coverage of practically all relevant issues. These texts are Chalmers’s (1999) What Is This Thing Called Science? and Card and Cover’s (1998) Philosophy of Science: The Central Issues.

2. To a considerable extent, current textbooks convey the picture of science as a finished product that has reached an endpoint in development. In stark contrast, the major message that comes from contemporary scholarship is that science has undergone considerable evolution with respect to its methods, which likely will continue. Thus, the greatest single idea that an instructor can impart to students is that a methodological principle accepted today may be rejected, modified, or supplanted by some new principle tomorrow. A major way to get this important idea across to students is to present them with historical material, detailing the methodological changes that have occurred in science. The best single source of information concerning this historical change is Laudan’s (1981) Science and Hypothesis: Historical Essays on Scientific Methodology. Laudan’s (1996) book, Beyond Positivism and Relativism: Theory, Method, and Evidence, discusses more recent developments in the methodology of science.

3. Having emphasized this historical material, instructors will find themselves in a position to demonstrate to students that scientists evaluate and decide on methodological rules as empirical propositions appear in Giere (1988), Laudan (1996), and Mayo (1996).

4. In covering hypothesis testing, instructors should make students aware that hypothesis testing is not a cut-and-dried affair. Essentially, what we are suggesting is that instructors should recognize and become aware of the Duhem–Quine thesis (see especially, Chalmers, 1999; Lakatos & Feyerabend, 1999), discussed earlier. Instructors should provide students with a number of detailed alternative courses of action to consider taking when a hypothesis is disconfirmed, for example, the advantages of introducing auxiliary hypotheses. The central idea that instructors should convey to students is that hypothesis testing is a great deal more complicated than it seems on the surface.

5. The instructor should provide students with concrete examples of how real, living scientists behaved under a variety of circumstances. For example, what have scientists done when a hypothesis has been confirmed? What have they done when a hypothesis has been disconfirmed? How have scientists reacted to other scientists claiming confirmation or disconfirmation of a hypothesis? The concrete examples provided can be either actual case histories gleaned from the historical record or idealizations of these case histories, provided that these idealizations clarify rather than distort the process. Ideally, the case histories provided will include examples from psychology as well as from other sciences. Case histories drawn from psychology occur in the latter chapters of Kuhn (1962) and Hanson (1958). Roberts (1998) supplied an interesting contemporary case history of scientific conduct in a research program concerned with the psychology of timing behavior. According to Roberts, the most interesting aspect of his findings “did not come from strong inference or hypothesis testing” (p. 101). Donovan et al. (1992) provided a variety of case histories drawn from other sciences.

6. Instructors should convey to students the importance of major research traditions as guiding forces in the construction of scientific theories. The importance of research traditions in science first was recognized in Kuhn’s (1962) monumental work, The Structure of Scientific Revolutions. Kuhn referred to these traditions as paradigms. Since Kuhn, much writing about research traditions has occurred, and their importance in giving rise to scientific theories is increasingly recognized. Lakatos (1970) and Laudan (1996) wrote extensively on research traditions within the context of the philosophy of science, with an emphasis on physics. More specifically relevant to psychology are articles and books indicating how and why psychologists came to accept or reject research traditions such as behaviorism and cognitive psychology (see Baars, 1986; Watson, 1913). Gholson and Barker (1985) provided excellent examples of how research traditions come to function in psychology.

7. Instructors should provide coverage of scientific methodologies other than hypothesis testing. The role of explanation in science is an outstanding example. As already indicated, Oreskes (1999) explained in detail how plate tectonics was originally proposed on the basis of already known data; that is, in terms of its explanatory capacity. Extensive and detailed treatments of several other important scientific theories that were accepted on the basis of explanation, rather than prediction are available to instructors. Finocchiaro


Research Methods Textbooks: An Objective Analysis

Sherri L. Jackson
Susan M. Lugo
Jacksonville University
Richard A. Griggs
University of Florida

Research methods is one of the more prevalent and frequently required courses in the undergraduate psychology curriculum. To aid teachers in the textbook selection process for this course, we compared 26 methods textbooks published between 1995 and 1999. We analyzed the texts’ demographic qualities, use of pedagogical aids and illustrative material, and topic coverage. The variability observed in these analyses should help teachers to find the text that best suits their preferences and needs.

The curriculum committee at the St. Mary’s Conference on Enhancing the Quality of Undergraduate Education in Psychology (Brewer et al., 1993) reported that research methods courses are especially important to undergraduate psychology education because they “foster analytical skills,” “encourage critical thinking,” and provide skills that can “enhance lifelong learning” (p. 173). Messer, Griggs, and Jackson (1999) found that 73% of the 518 degree options in their undergraduate curriculum survey required a research methods course. Similarly, Perlman and McCann (1999a), in their frequency count study of undergraduate psychology course listings, found that not only was research methods one of the most frequently listed courses but also that it had increased more than 10% since 1975. In their companion study on the structure of the undergraduate psychology curriculum, Perlman and McCann (1999b) found that other than the introductory, statistics, and capstone courses, research methods was the most frequently required course for psychology majors. Additionally, in his “psychological literacy” study of important terms for psychology education, Boneau (1990) found that 42 of the “top 100” terms were from research methods—statistics. Clearly, research methods courses and their content are essential to undergraduate education in psychology.

Given their importance, articles concerned with research methods have been a mainstay in Teaching of Psychology. Johnson (1995) provided an index for articles published in Teaching of Psychology between the years of 1974 and 1994. More articles were published on research methods (83) than any other topic. In a recent content analysis of Teaching of Psychology from 1991 through 1998, Griggs, Jackson, Christopher, and Marek (1999) found that articles on research methods averaged about 1 per issue. Surprisingly, however, not 1 article from 1974 through 1999 provided comparative data on research methods textbooks. The primary goal of this study was to fill that void. Using analytic techniques similar to those employed in recent studies of textbooks for other major psychology courses (e.g., Christopher, Griggs, &

References


Notes

1. I thank Patti Connor-Greene for her helpful suggestions on an earlier draft of this article.
2. Send correspondence and requests for specific psychometric course assignments to Cynthia Pury, Department of Psychology, Clemson University, 418 Brackett Hall, Clemson, SC 29634-1355; email: cpury@clemson.edu.
Hagans, 2000; Marek & Griggs, 2001), we compiled data on the objective features, content coverage, and pedagogical and illustration programs of research methods textbooks presently available for course adoption. These data should facilitate the textbook selection process for teachers of this important undergraduate course. In addition, these data will provide a comparative baseline for future studies of research methods texts.

Method

We obtained copies of all 26 full-length, undergraduate research methods textbooks published in the United States with copyrights between 1995 and 1999. References for the 26 texts appear in the Appendix.

First, we compiled basic demographic information for each textbook, including information on the number of textbook edition, authors, chapters, text pages, text pages plus back matter (e.g., references), and appendixes. We also calculated the average number of pages per chapter and noted the types of appendixes and type of text organization. We classified each text as organized from nonexperimental (descriptive, correlational, and quasi-experimental) to experimental methods, experimental to nonexperimental methods, or as mixed (coverage switched back and forth between the two types of methods).

We next inspected each textbook for its inclusion of seven pedagogical aids (see Table 2). We checked for use of chapter outlines, other chapter advance organizers (e.g., learning objectives), chapter and section summaries, glossaries of key terms, the use of bold and italic type styles, the use of verbal exercises, and the use of numeric exercises.

We also examined each textbook for the use of illustrative material. We counted the number of figures, tables, boxes, and formulas in each text. Because much of the statistical material is sometimes included in appendixes, we also included the number of formulas in the appendixes. In addition, we counted the percentage of formulas presented in verbal (e.g., $F = \frac{\text{systematic variance} + \text{error variance}}{\text{error variance}}$) format versus numeric or notational (e.g., $F = MS_t/MS_e$) format in both the text proper and appendixes.

We examined each of the tables of contents determined which areas received extended or full-chapter coverage. We included topics that received such coverage in at least one third of the texts in the content analysis. We included topics (e.g., applied research) not meeting this criterion in an “other” category. Although no text had more than 16 chapters and the mean number of chapters in the texts was 13.6, the number of content areas counted was 18 (including the “other” category) due to the variety of topic areas across the texts. We counted the number of pages in each text devoted to each of the 18 areas and converted to percentages based on the total number of pages in the text. For topics paired together in a chapter but considered as separate in the content analysis, we first counted the pages, or portions of pages (in ½-page increments), devoted to each topic, and then we proportionally allocated the chapter’s pages devoted to preview or summary materials among the relevant topics. Three content areas (APA style, ethics, and statistics), typically covered mainly in the text proper, were also frequently covered in appendixes. For these areas, we included both coverage in the text and appendixes in their content percentages.

Results and Discussion

The data for demographics, pedagogical aids, illustrative material, and content analyses appear in Tables 1, 2, 3, and 4, respectively. With respect to demographics, the texts represented a mix of new and old, with about half first or second editions and the other one half ranging from third to seventh editions. The texts varied in length from 11 to 16 chapters with a mean of 13.6 chapters. The average number of pages per text was 361.3 with a range of 183 to 548 pages. Most of the texts had appendixes ($M = 3.2$), mainly for statistical tables and tests and APA style information. Last, the majority ($M = 3.2$) of the texts were organized from nonexperimental to experimental methods.

Slightly more than half ($M = 3.3$) of the texts provided chapter outlines, but fewer than half ($M = 3.3$) used other chapter advance organizers (see Table 2). Most of the texts provided chapter summaries ($M = 3.2$), some type of glossary ($M = 3.2$), and chapter lists of key terms ($M = 3.2$). Whereas most used bold type style ($M = 3.3$), all used italic type style. All but 1 of the texts provided verbal exercises either at the end of each chapter or throughout the chapters, and 16 of the texts used numerical exercises of some sort.

With respect to illustrative material (see Table 3), most texts used both figures ($M = 53.3$) and tables ($M = 33$). The exact number of each, however, was highly variable across texts. Similarly, most texts used information boxes ($M = 13.8$), but how extensively they were used varied greatly. Last, most texts provided formulas in both verbal and numerical format, and yet the extent of use of each varied across texts.

Texts also varied greatly in their coverage of the 18 content areas (see Table 4). We found that the greatest variability occurred in coverage of statistics ($7\%$ to $36\%$). Other content areas with high variability of coverage included between-subjects experimental designs ($4\%$ to $28\%$), measurement ($0\%$ to $20\%$), the scientific method ($3\%$ to $16\%$), and...
Table 1. Demographic Data for the Individual Texts

<table>
<thead>
<tr>
<th>Textbook</th>
<th>Edition</th>
<th>No. of Authors</th>
<th>No. of Chapters</th>
<th>No. of Text Pages</th>
<th>M No. of Pages Per Chapter</th>
<th>Text Pages Plus Back Matter</th>
<th>No. of Appendixes</th>
<th>Types of Appendixes</th>
<th>Test Organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allen</td>
<td>1</td>
<td>1</td>
<td>14</td>
<td>253</td>
<td>18.1</td>
<td>267</td>
<td>0</td>
<td>a</td>
<td>n → e</td>
</tr>
<tr>
<td>Bordens &amp; Abbott</td>
<td>4</td>
<td>2</td>
<td>16</td>
<td>526</td>
<td>32.9</td>
<td>583</td>
<td>1</td>
<td>a</td>
<td>n → e</td>
</tr>
<tr>
<td>Christensen</td>
<td>7</td>
<td>1</td>
<td>15</td>
<td>548</td>
<td>39.3</td>
<td>590</td>
<td>2</td>
<td>a</td>
<td>n → e</td>
</tr>
<tr>
<td>Cozby</td>
<td>6</td>
<td>1</td>
<td>12</td>
<td>230</td>
<td>19.2</td>
<td>335</td>
<td>4</td>
<td>a, b, c, j</td>
<td>m</td>
</tr>
<tr>
<td>Elmes et al.</td>
<td>6</td>
<td>3</td>
<td>14</td>
<td>357</td>
<td>25.5</td>
<td>464</td>
<td>3</td>
<td>a, b</td>
<td>n → e</td>
</tr>
<tr>
<td>Goodwin</td>
<td>2</td>
<td>1</td>
<td>12</td>
<td>403</td>
<td>33.6</td>
<td>528</td>
<td>5</td>
<td>a, b, c, e, f</td>
<td>e → n</td>
</tr>
<tr>
<td>Graziano &amp; Raulin</td>
<td>3</td>
<td>2</td>
<td>15</td>
<td>363</td>
<td>24.2</td>
<td>464</td>
<td>6</td>
<td>a, b, c, e, g, i</td>
<td>n → e</td>
</tr>
<tr>
<td>Heiman</td>
<td>2</td>
<td>1</td>
<td>11</td>
<td>382</td>
<td>34.7</td>
<td>521</td>
<td>4</td>
<td>a, b, c, d</td>
<td>e → n</td>
</tr>
<tr>
<td>Herzog</td>
<td>1</td>
<td>1</td>
<td>15</td>
<td>291</td>
<td>19.4</td>
<td>378</td>
<td>2</td>
<td>a, c</td>
<td>n → e</td>
</tr>
<tr>
<td>Jones</td>
<td>1</td>
<td>1</td>
<td>11</td>
<td>267</td>
<td>24.2</td>
<td>356</td>
<td>3</td>
<td>a, c, d</td>
<td>n → e</td>
</tr>
<tr>
<td>Leary</td>
<td>2</td>
<td>1</td>
<td>16</td>
<td>371</td>
<td>23.2</td>
<td>422</td>
<td>2</td>
<td>a, b</td>
<td>n → e</td>
</tr>
<tr>
<td>Liebert &amp; Liebert</td>
<td>4</td>
<td>2</td>
<td>11</td>
<td>234</td>
<td>23.1</td>
<td>282</td>
<td>2</td>
<td>b, f</td>
<td>n → e</td>
</tr>
<tr>
<td>Loos</td>
<td>1</td>
<td>1</td>
<td>16</td>
<td>370</td>
<td>23.1</td>
<td>401</td>
<td>1</td>
<td>j</td>
<td>n → e</td>
</tr>
<tr>
<td>McBurney</td>
<td>4</td>
<td>1</td>
<td>14</td>
<td>356</td>
<td>25.4</td>
<td>434</td>
<td>7</td>
<td>a, b, e, j</td>
<td>n → e</td>
</tr>
<tr>
<td>McGuigan</td>
<td>7</td>
<td>1</td>
<td>12</td>
<td>379</td>
<td>31.6</td>
<td>480</td>
<td>6</td>
<td>a, c, e, g, h, j</td>
<td>m</td>
</tr>
<tr>
<td>Mitchell &amp; Jolley</td>
<td>3</td>
<td>2</td>
<td>14</td>
<td>503</td>
<td>35.9</td>
<td>624</td>
<td>7</td>
<td>a, d, h, i, j</td>
<td>e → n</td>
</tr>
<tr>
<td>Myers &amp; Hanson</td>
<td>4</td>
<td>2</td>
<td>16</td>
<td>477</td>
<td>29.8</td>
<td>567</td>
<td>3</td>
<td>a, b, d</td>
<td>n → e</td>
</tr>
<tr>
<td>Nation</td>
<td>1</td>
<td>1</td>
<td>14</td>
<td>386</td>
<td>27.6</td>
<td>430</td>
<td>1</td>
<td>d</td>
<td>e → n</td>
</tr>
<tr>
<td>Ray</td>
<td>5</td>
<td>1</td>
<td>16</td>
<td>413</td>
<td>25.8</td>
<td>476</td>
<td>4</td>
<td>a, d, j</td>
<td>e → n</td>
</tr>
<tr>
<td>Rosnow &amp; Rosenthal</td>
<td>3</td>
<td>2</td>
<td>15</td>
<td>342</td>
<td>22.8</td>
<td>475</td>
<td>3</td>
<td>a, b, c, d</td>
<td>m</td>
</tr>
<tr>
<td>Salkind</td>
<td>3</td>
<td>1</td>
<td>13</td>
<td>316</td>
<td>24.3</td>
<td>380</td>
<td>3</td>
<td>e, g, j</td>
<td>n → e</td>
</tr>
<tr>
<td>Schweigert</td>
<td>1</td>
<td>1</td>
<td>12</td>
<td>183</td>
<td>15.3</td>
<td>275</td>
<td>3</td>
<td>a, b, c</td>
<td>e → n</td>
</tr>
<tr>
<td>Shaughnessy &amp; Zechmeister</td>
<td>4</td>
<td>2</td>
<td>10</td>
<td>379</td>
<td>37.9</td>
<td>520</td>
<td>3</td>
<td>b, c, d, f</td>
<td>n → e</td>
</tr>
<tr>
<td>Smith &amp; Davis</td>
<td>1</td>
<td>2</td>
<td>12</td>
<td>445</td>
<td>37.1</td>
<td>486</td>
<td>2</td>
<td>a, e</td>
<td>n → e</td>
</tr>
<tr>
<td>Stangor</td>
<td>1</td>
<td>1</td>
<td>14</td>
<td>268</td>
<td>19.1</td>
<td>422</td>
<td>5</td>
<td>a, b, c, d</td>
<td>n → e</td>
</tr>
<tr>
<td>Vadum &amp; Rankin</td>
<td>1</td>
<td>2</td>
<td>13</td>
<td>352</td>
<td>27.1</td>
<td>419</td>
<td>2</td>
<td>d, j</td>
<td>m</td>
</tr>
<tr>
<td>M</td>
<td>3.0</td>
<td>1.4</td>
<td>13.6</td>
<td>361.3</td>
<td>26.9</td>
<td>445.5</td>
<td>3.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. The full references for the textbooks are in the Appendix.

*a* Types of appendixes: (a) statistical tables, (b) statistical tests, (c) APA style, (d) sample APA paper, (e) answer keys, (f) developing surveys, (g) computer use, (h) ethics, (i) library use/research, and (j) other.

*b* Texts were classified as going from nonexperimental methods to experimental methods (n → e), as going from experimental methods to nonexperimental methods (e → n), or as having a mixed organization—going back and forth between the two types of methods (m). Nonexperimental methods = descriptive, correlational, and quasi-experimental methods.
### Table 2. Pedagogical Aids in Individual Texts

<table>
<thead>
<tr>
<th>Textbook</th>
<th>Chapter Outline</th>
<th>Chapter Advance Organizers</th>
<th>Type of Summary</th>
<th>Type of Glossary</th>
<th>Listed at End of Chapter</th>
<th>M No. per Chapter</th>
<th>Type Style</th>
<th>Verbal Exercises</th>
<th>Numeric Exercises</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allen</td>
<td>Chapter</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>465</td>
<td>33.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bordens &amp; Abbott</td>
<td>Chapter</td>
<td></td>
<td>Book</td>
<td></td>
<td></td>
<td>278</td>
<td>17.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Christenson</td>
<td>O</td>
<td>Chapter</td>
<td>Running</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cozby</td>
<td></td>
<td>Book</td>
<td></td>
<td></td>
<td></td>
<td>161</td>
<td>13.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elmes et al.</td>
<td>S</td>
<td>Section &amp; chapter</td>
<td>Book</td>
<td></td>
<td></td>
<td>261</td>
<td>18.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Goodwin</td>
<td></td>
<td>S</td>
<td>Book</td>
<td></td>
<td></td>
<td>245</td>
<td>20.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Graziano &amp; Raulin</td>
<td></td>
<td>Chapter</td>
<td>Book</td>
<td></td>
<td></td>
<td>443</td>
<td>29.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heiman</td>
<td>O</td>
<td>Section &amp; chapter</td>
<td>Book</td>
<td></td>
<td></td>
<td>317</td>
<td>28.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Herzog</td>
<td></td>
<td>Chapter</td>
<td>Book</td>
<td></td>
<td></td>
<td>297</td>
<td>19.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jones</td>
<td></td>
<td>Chapter</td>
<td>Book</td>
<td></td>
<td></td>
<td>228</td>
<td>20.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leary</td>
<td>Chapter</td>
<td>Book</td>
<td></td>
<td></td>
<td></td>
<td>297</td>
<td>18.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liebert &amp; Liebert</td>
<td></td>
<td></td>
<td>Book</td>
<td></td>
<td></td>
<td>299</td>
<td>27.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loos</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>McBurney</td>
<td>S</td>
<td>Chapter</td>
<td>Running</td>
<td></td>
<td></td>
<td>177</td>
<td>12.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>McGuigan</td>
<td>O</td>
<td>Section &amp; chapter</td>
<td>Book</td>
<td></td>
<td></td>
<td>216</td>
<td>18.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mitchell &amp; Jolley</td>
<td>S</td>
<td>Chapter</td>
<td>Chapter &amp; book</td>
<td></td>
<td></td>
<td>216</td>
<td>15.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Myers &amp; Hanson</td>
<td></td>
<td>Chapter</td>
<td>Chapter &amp; book</td>
<td></td>
<td></td>
<td>265</td>
<td>16.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nation</td>
<td></td>
<td></td>
<td>Running</td>
<td></td>
<td></td>
<td>186</td>
<td>13.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ray</td>
<td></td>
<td>Chapter</td>
<td>Book</td>
<td></td>
<td></td>
<td>96</td>
<td>6.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rosnow &amp; Rosenthal</td>
<td>Q</td>
<td>Chapter</td>
<td>Book</td>
<td></td>
<td></td>
<td>465</td>
<td>31.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Salkind</td>
<td>S</td>
<td></td>
<td>Book</td>
<td></td>
<td></td>
<td>179</td>
<td>13.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schweigert</td>
<td></td>
<td>Chapter</td>
<td>Book</td>
<td></td>
<td></td>
<td>262</td>
<td>21.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shaughnessy &amp; Zechmeister</td>
<td>S</td>
<td>Chapter</td>
<td>Book</td>
<td></td>
<td></td>
<td>136</td>
<td>13.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smith &amp; Davis</td>
<td></td>
<td>Section &amp; chapter</td>
<td>Running</td>
<td></td>
<td></td>
<td>310</td>
<td>25.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stangor</td>
<td>Q</td>
<td>Chapter</td>
<td>Book</td>
<td></td>
<td></td>
<td>328</td>
<td>23.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vadum &amp; Rankin</td>
<td></td>
<td></td>
<td>Book</td>
<td></td>
<td></td>
<td>345</td>
<td>26.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note.** The full references for the textbooks are in the Appendix. A bullet (*) or an entry in a cell indicates the presence of an aid. O = listed objectives preceding chapter; S = summary preceding chapter; Q = preview questions.
Table 3. Illustrative Material in Individual Texts

<table>
<thead>
<tr>
<th>Textbook</th>
<th>Total Figures</th>
<th>Total Tables</th>
<th>Total Boxes</th>
<th>% in Verbal Format</th>
<th>% in Numerical Format</th>
<th>Formulas in Text</th>
<th>% in Verbal Format</th>
<th>% in Numerical Format</th>
<th>Formulas in Appendixes&lt;sup&gt;a&lt;/sup&gt;</th>
<th>% in Verbal Format</th>
<th>% in Numerical Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allen</td>
<td>25</td>
<td>11</td>
<td>33</td>
<td>1</td>
<td>100</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bordens &amp; Abbott</td>
<td>130</td>
<td>49</td>
<td>0</td>
<td>23</td>
<td>4</td>
<td>96</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Christensen</td>
<td>81</td>
<td>28</td>
<td>11</td>
<td>33</td>
<td>24</td>
<td>76</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cozby</td>
<td>33</td>
<td>9</td>
<td>1</td>
<td>5</td>
<td>20</td>
<td>80</td>
<td>18</td>
<td>6</td>
<td>94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elmes et al.</td>
<td>49</td>
<td>18</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>100</td>
<td>21</td>
<td>5</td>
<td>95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Goodwin</td>
<td>97</td>
<td>25</td>
<td>33</td>
<td>7</td>
<td>14</td>
<td>86</td>
<td>43</td>
<td>5</td>
<td>95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Graziano &amp; Raulin</td>
<td>42</td>
<td>57</td>
<td>15</td>
<td>12</td>
<td>67</td>
<td>33</td>
<td>38</td>
<td>0</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heiman</td>
<td>33</td>
<td>58</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>68</td>
<td>6</td>
<td>94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Herzog</td>
<td>19</td>
<td>38</td>
<td>0</td>
<td>23</td>
<td>4</td>
<td>96</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jones</td>
<td>28</td>
<td>18</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leary</td>
<td>46</td>
<td>15</td>
<td>0</td>
<td>30</td>
<td>20</td>
<td>80</td>
<td>17</td>
<td>0</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liebert &amp; Liebert</td>
<td>48</td>
<td>30</td>
<td>11</td>
<td>2</td>
<td>50</td>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loos</td>
<td>12</td>
<td>46</td>
<td>0</td>
<td>15</td>
<td>20</td>
<td>80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>McBurney</td>
<td>71</td>
<td>45</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>100</td>
<td>30</td>
<td>20</td>
<td>80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>McGuigan</td>
<td>71</td>
<td>52</td>
<td>0</td>
<td>40</td>
<td>18</td>
<td>82</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mitchell &amp; Jolley</td>
<td>47</td>
<td>87</td>
<td>18</td>
<td>15</td>
<td>47</td>
<td>53</td>
<td>1</td>
<td>0</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Myers &amp; Hanson</td>
<td>77</td>
<td>38</td>
<td>39</td>
<td>36</td>
<td>11</td>
<td>89</td>
<td>6</td>
<td>33</td>
<td>67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nation</td>
<td>103</td>
<td>3</td>
<td>22</td>
<td>12</td>
<td>42</td>
<td>58</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ray</td>
<td>62</td>
<td>15</td>
<td>24</td>
<td>24</td>
<td>58</td>
<td>42</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rosnow &amp; Rosenthal</td>
<td>27</td>
<td>58</td>
<td>103</td>
<td>58</td>
<td>16</td>
<td>84</td>
<td>9</td>
<td>0</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Salkind</td>
<td>46</td>
<td>45</td>
<td>0</td>
<td>12</td>
<td>8</td>
<td>92</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schweigert</td>
<td>26</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>100</td>
<td>0</td>
<td>30</td>
<td>0</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shaughnessy &amp; Zechmeister</td>
<td>46</td>
<td>23</td>
<td>3</td>
<td>3</td>
<td>100</td>
<td>0</td>
<td>26</td>
<td>12</td>
<td>88</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smith &amp; Davis</td>
<td>91</td>
<td>48</td>
<td>0</td>
<td>12</td>
<td>75</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stangor</td>
<td>32</td>
<td>22</td>
<td>0</td>
<td>12</td>
<td>50</td>
<td>50</td>
<td>40</td>
<td>7</td>
<td>93</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vadum &amp; Rankin</td>
<td>44</td>
<td>16</td>
<td>29</td>
<td>7</td>
<td>0</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;i&gt;M&lt;/i&gt;</td>
<td>53.3</td>
<td>33.0</td>
<td>13.8</td>
<td>15.1</td>
<td></td>
<td>24.8&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>Note.</sup> The full references for the textbooks are in the Appendix.

<sup>a</sup>A blank cell indicates no statistical appendix. The Liebert and Liebert text did have a statistical appendix, but did not use formulas.

<sup>b</sup>Mean is based on only those texts with statistical appendixes.
| Textbook          | Scientific Method | Developing Ideas | Ethics | APA Style | Statistics | Measurement | Observational | Survey | Correlational | Quasi-Experimental | Between Subject | Within Subject | Factorial | Single Subject | Field Studies | Archival Studies | Generalizing | Other |
|-------------------|-------------------|------------------|--------|-----------|------------|-------------|---------------|--------|---------------|---------------------|----------------|----------------|-----------|---------------|---------------|----------------|---------------|---------|-------|
| Allen             | 8                 | 8                | 5      | 13        | 15         | 12          | 5              | 2             | 4             | 6                   | 10             | 0              | 0         | 0             | 0              | 1               | 4            | 7      |
| Bordens & Abbott  | 12                | 5                | 7      | 14        | 21         | 6           | 5              | 7              | 2             | 3                   | 4              | 4              | 3         | 7             | 0              | 0               | 0            | 0      |
| Christensen       | 8                 | 5                | 12     | 9         | 7          | 5           | 3              | 2              | 1             | 4                   | 28             | 3              | 2         | 7             | 0              | 0               | 4            | 0      |
| Cozby             | 5                 | 8                | 5      | 17        | 26         | 11          | 6              | 4              | 0             | 3                   | 15             | 3              | 7         | 2             | 0              | 1               | 6            | 0      |
| Elmes et al.      | 11                | 11               | 6      | 10        | 13         | 7           | 7              | 1              | 5             | 6                   | 10             | 3              | 7         | 5             | 0              | 0               | 8            | 2      |
| Goodwin           | 7                 | 9                | 8      | 7         | 10         | 6           | 4              | 2              | 8             | 8                   | 17             | 6              | 8         | 9             | 0              | 1               | 1            | 0      |
| Graziano & Raulin | 15                | 5                | 2      | 2         | 11         | 7           | 6              | 3              | 3             | 6                   | 15             | 5              | 7         | 3             | 0              | 0               | 0            | 10     |
| Heiman            | 10                | 7                | 7      | 11        | 18         | 4           | 3              | 4              | 7             | 6                   | 11             | 3              | 11        | 3             | 4              | 0               | 1            | 10     |
| Herzog            | 12                | 0                | 6      | 10        | 24         | 10          | 6              | 8              | 2             | 0                   | 14             | 0              | 8         | 0             | 0              | 0               | 5            | 4      |
| Jones             | 9                 | 9                | 7      | 13        | 16         | 1           | 2              | 5              | 9             | 8                   | 22             | 1              | 8         | 0             | 0              | 1               | 0            | 0      |
| Leary             | 8                 | 0                | 6      | 10        | 24         | 10          | 3              | 3              | 3             | 10                  | 6              | 9              | 1         | 5             | 6              | 0               | 0            | 1      |
| Liebert & Liebert | 8                 | 5                | 5      | 7         | 12         | 10          | 2              | 0              | 9             | 3                   | 18             | 0              | 10        | 8             | 0              | 1               | 3            | 0      |
| Loos              | 10                | 9                | 3      | 5         | 21         | 15          | 5              | 0              | 3             | 0                   | 12             | 6              | 5         | 0             | 0              | 0               | 0            | 5      |
| McBurney          | 9                 | 2                | 6      | 12        | 16         | 7           | 6              | 8              | 0             | 8                   | 12             | 4              | 7         | 6             | 0              | 1               | 1            | 4      |
| McGuigan          | 12                | 11               | 3      | 6         | 24         | 0           | 0              | 0              | 3             | 6                   | 14             | 5              | 5         | 9             | 0              | 0               | 11           | 0      |
| Mitchell & Jolley | 8                 | 11               | 14     | 10        | 14         | 12          | 3              | 7              | 0             | 5                   | 16             | 7              | 9         | 4             | 0              | 1               | 0            | 0      |
| Myers & Hanson    | 5                 | 5                | 5      | 15        | 22         | 7           | 5              | 2              | 3             | 3                   | 16             | 5              | 4         | 4             | 0              | 0               | 5            | 1      |
| Nation            | 8                 | 8                | 8      | 10        | 11         | 11          | 2              | 2              | 1             | 3                   | 22             | 5              | 5         | 2             | 0              | 0               | 1            | 2      |
| Ray               | 14                | 8                | 8      | 11        | 15         | 12          | 2              | 6              | 1             | 3                   | 4              | 5              | 6         | 5             | 0              | 0               | 0            | 7      |
| Rosnow & Rosenthal| 7                 | 10               | 7      | 12        | 36         | 6           | 4              | 16             | 2             | 4                   | 7              | 0              | 1         | 1             | 1              | 1               | 1            | 0      |
| Salkind           | 9                 | 15               | 4      | 13        | 17         | 20          | 1              | 2              | 2             | 3                   | 12             | 0              | 0         | 0             | 0              | 0               | 1            | 9      |
| Schweigert        | 11                | 1                | 10     | 10        | 23         | 2           | 9              | 10             | 0             | 7                   | 8              | 8              | 8         | 9             | 0              | 5               | 1            | 0      |
| Shaughnessy & Zechmeister | 9 | 1 | 10 | 12 | 20 | 1 | 9 | 7 | 4 | 9 | 8 | 6 | 7 | 9 | 0 | 8 | 1 | 2 |
| Smith & Davis     | 3                 | 7                | 4      | 12        | 23         | 0           | 3              | 4              | 4             | 1                   | 3              | 23             | 2         | 5             | 3              | 0               | 0            | 4      |
| Stangor           | 7                 | 5                | 8      | 15        | 30         | 14          | 4              | 3              | 5             | 6                   | 13             | 2              | 8         | 1             | 0              | 1               | 5            | 2      |
| Vadum & Rankin    | 16                | 8                | 9      | 16        | 11         | 8           | 4              | 4              | 4             | 3                   | 3              | 10             | 1         | 6             | 7              | 0               | 1            | 0      |
|M                  | 9.1               | 6.6              | 7.1    | 10.9      | 18.4       | 7.9          | 4.2            | 3.3            | 4.4           | 4.7                  | 13.5           | 3.3            | 5.9        | 4.3           | 0.4            | 0.9             | 2.5         | 2.6   |

Note. The full references for the textbooks are in the Appendix. Rows may not add to exactly 100% due to rounding.

*The percentages of text devoted to this topic include coverage in appendixes. Both nonexperimental and experimental single-subject designs are included in this calculation if identified as such in a text.

This category includes all topics not receiving chapter or extended coverage in at least one-third of the texts (e.g., applied research).
developing ideas (0% to 15%). On the average, experimental methods received the most coverage (23%), with statistics (18%) and nonexperimental methods (17%) following closely.

In conclusion, the primary goal in conducting this study was to provide teachers with comparative, objective data on research methods texts to assist them in the text selection process. These data should help teachers to narrow the set of texts that they can more thoroughly evaluate for selection. A secondary motive in conducting the research was to provide archival baseline information on research methods texts. Given the detailed nature of the analyses, both goals have been met. There is now a set of comparative data on undergraduate methods texts, and these data should facilitate teachers’ search for a text that meshes with their individual course objectives.

References


Appendix

Research Methods Texts


Notes

1. We want to thank all of the publishers who supplied us with the texts used in this study and Randolph Smith and three anonymous reviewers for their comments on an earlier version of this article.
2. Send correspondence to Sherri L. Jackson, Department of Psychology, Jacksonville University, 2800 University Boulevard North, Jacksonville, FL 32211–3394; e-mail: sjackson@ju.edu.
In this article, we illustrate a simple field experiment that facilitates student understanding of several important methodological issues, including random assignment, equalizing the strength of manipulations, and experimenter bias. The basic hypothesis of the field experiment is that people respond to smiles with smiles but fail to reciprocate frowns. Several classroom replications have demonstrated that this effect is reliable, although the processes underlying the findings are still unknown and consequently generate useful theoretical discussions by students. An evaluation of the demonstration’s effectiveness showed that students increased their understanding of experimentation and attributed much of their learning to the field experiment.

The use of class demonstrations to help students learn important psychological concepts is well documented (Costanzo & Archer, 1991; Goldstein, Hopkins, & Strube, 1994; Hom, 1994; Jones, 1994). Demonstrations that instruct students on important experimental concepts are also fairly common (Lutsky, 1993; Marnie, 1994; Miserandino, 1992). We have not, however, encountered an experimental demonstration that addressed the scope of issues reflected in the field experiment presented in this article.

Our students conducted a replication of Hinsz and Tomhave’s (1991) field experiment to learn a variety of research methods objectives: (a) the importance of random assignment, (b) issues related to the strength of experimental manipulations, and (c) how to avoid experimenter bias. This class project allows students to conduct an enjoyable, unobtrusive field experiment from hypothesis derivation through American Psychological Association-style report writing. Advantages to this demonstration are (a) its ease of completion, (b) no materials are necessary (other than the students’ faces), and (c) the results are highly reliable.

Method

Providing the Framework

We first provided an explanation of the experiment and encouraged students to read Hinsz and Tomhave (1991). Hinsz and Tomhave tested a contagion hypothesis that people reciprocate facial expressions displayed toward them. Their students went to public places and presented a smile, a neutral expression, or a frown to unsuspecting passersby. They found that whereas half (52.6%) of the people returned smiles, few (4.6%) returned frowns. We then provided specific instructions about the experimental methods as follows.

Random Assignment

We told students to write down each of the stimulus expressions (smile, neutral, frown) five times on 15 separate pieces of paper, put them in a hat, and draw them one at a time to determine the random presentation of the stimuli. When discussing the results of the experiment, students frequently suggested other plausible influences for the behaviors of the passersby (e.g., the weather, the location). However, when reminded about random assignment of participants to conditions, students came to realize that although such influences could impact some individual responses, randomly assigning participants eliminated the concern that the data resulted from such extraneous variables.

Equalizing Manipulation Strength

A combination of operational definitions, mental imagery, and practice helped students create consistent facial expressions. For example, the operational definition for a smile (frown) was as follows: “The eyes wide open (focused), the forehead is not creased (creased), and the corners of the mouth are pulled back and turned up (turned down).” Likewise, students expressed their feelings through facial expressions after imagining “that they were with a good friend that they enjoyed being with very much” (“that they were with a close friend and that something very bad happened to that friend”). Finally, they devoted about 30 min to practice in pairs creating the three facial stimuli until they believed they were reliably producing the facial expressions. Frowns proved
to be more difficult to master than smiles and required most of the practice time. This exercise facilitated student learning about how unequal strength of the manipulations can influence the results of an experiment. More people responding with smiles than frowns is potentially less interesting if smiles were simply more strongly manipulated than frowns. As a result, students closely examined the issue of manipulation strength when conducting this experiment. There was no check for the reliability of stimulus expressions included in these demonstrations. Issues pertaining to reliability and the possible variance in manipulation strength could be tested in future class efforts by including manipulation checks.

Reducing Experimenter Bias

To reduce experimenter bias, students practiced working with another pair (who acted as passersby) to perfect their coding of the response expressions of the passersby. “Displayers,” who had exclusive knowledge and were in sole possession of the random orders of their stimulus presentations, presented their facial expressions to passersby and noted if the passersby noticed the expression. “Recorders” walked about 1.5 m behind the displayers. After making eye contact with the passersby, the displayer unobtrusively signaled the recorder with a behind-the-back gesture to note that the passersby should be counted as a participant. The recorder then recorded the passersby's facial expression. Through this process, students learned about the importance of using experimenters who are blind to the stimulus presentation condition. They saw for themselves that if the data recorders knew the order of presentation, they could be biased in their interpretation of the responses of the passersby.

Theoretical Discussions

This experiment also stimulated students’ consideration of various theories to account for the results. Several plausible theoretical explanations for the findings emerged through class discussion, including the norm of reciprocity (as suggested by Hinsz & Tomhave, 1991), attribution theory (Kunda, 1990), and evolutionary psychology theory (Buss, 1994; Hansen & Hansen, 1988).

Through a discussion of these theoretical explanations of the findings, the students evaluated how opposing theories help generate further research. These theories may lead to variations on the specific methodology presented in this article to discern which explanation best accounts for the findings.

Procedure

The student experimenter pairs decided on a time and location outside of class to perform the experiment. Students chose areas with high pedestrian traffic such as malls, parks, or the campus student union. Experimenters did not debrief passersby because the manipulations were unobtrusive and no physical or mental risks were associated with the procedure. On average, each pair collected their data in 1 hr. The instructor collected the data sheets and analyzed the combined data for dissemination in the next class meeting.

Testing the Effectiveness of the Demonstration

To test the effectiveness of this demonstration on learning, students in one of the replications of this experiment completed a pretest and a posttest measuring the concepts discussed earlier: random assignment, manipulation strength, and experimenter bias. Students completed a six-question pretest for extra credit during their midterm exam, which was held during the class meeting immediately preceding this project.

The first three questions asked the students to define each of the methodological concepts. The students then read the primary hypothesis and a brief summary of the procedure and answered these three questions:

1. Regarding the experiment described, what factors do you think would be controlled for by using random assignment? (List all factors that apply.)
2. When developing a study to test the hypothesis explained, you find that it is easier to smile than to frown. How would you, as an experimenter, ensure that both expressions are performed equally well during the experiment? (List all different methods you would use to do this.)
3. Concerning the experiment described, how might the experimenter's knowledge of the hypothesis affect the results?

Students completed a posttest consisting of the same questions when they finished the project.

Results and Discussion

Replicability of the Demonstration

We used a log-linear analysis (because the data were categorical; for a discussion of this analysis, see Stevens, 1992) to test the hypothesis that passersby smile at smiles more than they frown at frowns.1 A significant main effect (shown in Figure 1) replicated previous findings that passersby were more likely to return a smile with a smile (55.4%) than they were to return a frown with a frown (8.5%). The data supported the main effect of the manipulation (55.4% vs. 8.5%), LR(4, N = 1,546) = 139.73, p < .0001. No interaction emerged between the dependent variable of respondent and sex of experimenter to test for the sex effects predicted by Hinsz and Tomhave (1991). Similar results emerged from these analyses and demonstrated reliability between replications. If anyone is interested in these data, please contact the first author. Because the purpose of this article is to demonstrate the effectiveness of this project as a teaching tool and not as a research report, we present only the main effect of stimulus expression from this design. However, for those interested, Hinsz and Tomhave (1991) previously described the sex effects that we replicated.

---

1We also conducted a four-factor design on the data including sex of respondent and sex of experimenter to test for the sex effects predicted by Hinsz and Tomhave (1991). Similar results emerged from these analyses and demonstrated reliability between replications. If anyone is interested in these data, please contact the first author.
tween replication number and stimulus expression, \( LR^2 \) \((12, N = 1,546) = 5.74, \text{ ns}, \) suggesting that the main effect was stable across the four replications.

We recently put this class project to a cross-cultural test in which 144 social psychology students in Sydney, Australia, replicated this exercise. The results remained remarkably stable, although the Australian sample tended to return frowns more frequently: 51.1\% of the respondents responded to a smile with a smile, whereas 16.9\% of the respondents responded to a frown with a frown.

**Effectiveness of Demonstration**

Eleven students completed both the pretest and posttest in the most recent U.S. replication. Two independent raters evaluated each student’s answer to all six questions. For the definition questions, the raters scored each answer based on a scale ranging from 1 (incorrect response) to 4 (correct response). For the short-answer questions, the raters counted each correct response to a given answer. The median reliability was .74. The mean of the two raters’ scores represented the students’ degree of correctness for each question. The sum of the first three questions comprised the students’ definition scores and the sum of the last three questions comprised the students’ short-answer scores. We conducted a one-way, repeated measures ANOVA on each of these scores. Students’ definition scores did not differ statistically between the posttest (\( M = 7.8, SD = 2.1 \)) and the pretest (\( M = 7.2, SD = 1.9 \)), \( F(1, 10) < 1, \text{ ns.} \) However, students achieved higher short-answer scores in the posttest (\( M = 5.4, SD = 2.6 \)) than the pretest (\( M = 4.2, SD = 2.2 \)), \( F(1, 10) = 4.98, p < .05. \)

These data suggest that this demonstration improved students’ ability to think critically about the hypothesis and methodology. Students’ prior exposure to the definitions in lectures or the text may be a reason why no improvement emerged in the definition scores.

Students also reported the extent to which the class demonstration influenced their learning by listing the percentage of their learning they attributed to reading the text, the lectures, the class project, writing assignments, outside readings, or some other source. Students completed these ratings, along with the general class evaluations, for five questions: general experimental issues; understanding of issues related to random assignment, manipulation strength, and experimenter bias; and recognition of alternative explanations of an experiment’s results.

For all five questions, students attributed much of their learning to lectures and the assigned texts, but they attributed a substantial portion to the class project ( Median = 17.6\%). The percentage attributed to lectures, texts, and the class project for each of the five questions appears in Table 1.

The 144 Australian students reported what they learned from this class demonstration in a free-response format (see Table 2). Many stated that they learned about the expected issues (especially random assignment and experimenter bias). Some mentioned issues that we had not anticipated (e.g., the meaning of replications and the generalizability of a study).

The final illustration of the effectiveness of this project is the enjoyment many students reported. Many remarked about the positive experience while conducting the experiment and specifically commented that the experiment was
easy to perform and easy to understand and that it instilled a
greater interest in observing and understanding their social
world. Laughter and outward signs of enthusiasm character-
ized the students’ behavior during practice and when they re-
called specific instances during the experiment.

In summary, this project is particularly useful in an experi-
mental class for a number of reasons. First, the instructor can
rely on replicating the hypothesized results. Second, the ex-
periment requires no special equipment or facilities. Finally,
most students commented that the experiment is easy and
fun to perform.

References

223–226.
the beholder”: A classroom demonstration of observer bias.
crowd: An anger superiority effect. Journal of Personality and Social
Psychology, 54, 917–924.
Hinsz, V. B., & Tomhave, J. A. (1991). Smile and (half) the world
smiles with you, frowned and you frowned alone. Personality and Social

Notes

1. Some of the results were presented at the 68th annual meeting of
the Midwestern Psychological Association, Chicago, May 1996.
2. We thank Frank Bernieri and Scott Madey for allowing us to
collect information about this class demonstration in their
courses.
3. Send correspondence to Jon E. Grahe, Department of Psychology,
Monmouth College, Monmouth, IL 61462; e-mail: jgrahe@monm.edu.
Research Methods with a Smile: A Gender Difference Exercise that Teaches Methodology
Angela Lipsitz

Teaching of Psychology 2000 27: 111
DOI: 10.1207/S15328023TOP2702_07

The online version of this article can be found at:
http://top.sagepub.com/content/27/2/111

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Society for the Teaching of Psychology

Additional services and information for Teaching of Psychology can be found at:
Email Alerts: http://top.sagepub.com/cgi/alerts
Subscriptions: http://top.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Apr 1, 2000

What is This?
Research Methods With a Smile: A Gender Difference Exercise That Teaches Methodology

Angela Lipsitz
Northern Kentucky University

In this activity, I provide a way to collect data on a gender difference—smiling in yearbook photos—without leaving the classroom. I also provide an effective procedure for reviewing or introducing a number of methodological concepts. Two sections of Psychology of Gender students (N = 45) rated the activity high on achievement of purpose and enjoyability, and 95% recommended using the exercise in future classes.

Over the years, instructors have developed a variety of classroom exercises and computer simulations for teaching research methodology. These techniques include teaching hypothesis testing by having students prove their instructor is not a psychic (Bates, 1991), teaching random assignment by designing a basketball experiment (Watson, 1990), teaching observer bias by falsely stating that a research subject has consumed alcohol (Goldstein, Hopkins, & Strube, 1994), and teaching statistical interaction by using variations on a Stroop task (Vernoy, 1994).

I have a classroom exercise that I use to review null and alternative hypothesis testing, operational definitions, interpretation of results, and external validity. Although created for a gender class, the activity could be a useful tool for research methods, social, or developmental psychology as well. A number of researchers (Briton & Hall, 1995; Hall, 1984, 1987; Hinsz & Tomhave, 1991; Kolaric & Galambos, 1995) have found that, at least among adults, women smile more than men. As this finding is reliably observed in high school and college yearbooks (Deaux, 1976; Ragan, 1982), I have students conduct a gender difference study using individual yearbook photos for the data collection.
Procedure

Collecting the Data

On a designated day I have students bring a yearbook from a coed school to class. I usually bring in an extra in case someone forgets. (Although I vainly bring in the yearbook in which I look my slimmest and most attractive, instructors with higher self-esteem might want to bring in an annual in which they look humorous or grossly dated.) I tell students that they will be examining gender differences in smiling, using the individual class sections of their annuals for the data collection. I indicate that they will be testing the null hypothesis that there are no gender differences in smiling, and I ask them to propose an alternative hypothesis. Usually, students hypothesize that female students will smile more. I also explain that they will need an operational definition of smiling, and I ask them to look casually through the photos in their yearbooks for some possibilities. I write several proposals on the board and then have students consider these proposals as they look through their yearbooks again. Students eventually reach a consensus on which definition will work best. Typically, they choose definitions such as “teeth showing,” “visible cheek lines,” or “upturned mouth.” I explain that it is unlikely that any definition will correctly capture all cases. For example, if smiling is defined as teeth showing, there will probably be a few individuals who show teeth but do not look happy and a few who look quite happy even though their mouths are completely closed. I emphasize that it is important to employ the operational definition rigorously, despite these exceptions.

Although students can do this activity individually, they find it easier and more enjoyable when working in pairs. Thus, I have pairs of students record smiling data from one class (sophomore, junior, etc.) from each of their yearbooks. I ask them first to make private judgments, then to share their judgments and, if necessary, to resolve any differences. Students tally smiling/not smiling separately for each gender. To minimize preexisting bias, I tell students not to record data from their own classes. To minimize frustration, I ask students to avoid using classes with extremely small photos. So that the student pairs finish in approximately the same amount of time, I have two other stipulations: If the class size is over 400, students should record data for every second or third page, and if the class size is very small, students should record data for more than two classes.

As students finish, I have them record the following data on the board or an overhead transparency: yearbook publication date, school type (e.g., high school, college), class (e.g., junior, senior), total class size, coders’ genders, percentage of male students smiling, and percentage of female students smiling. Results routinely show that more female students than male students smile in their yearbook class photos. Averaging across yearbooks, students typically find about 90% of the female students smiling and about 67% of the male students smiling. Percentages will fluctuate depending on the operational definition employed, and, of course, results for individual yearbooks may be quite variable. The only cases in which students have found more male students smiling have been in elementary school yearbooks. This latter result is consistent with Hall’s (1984, 1987) meta-analysis of children’s smiling.

I can usually fit the setup, data collection, and discussion of methodological issues (see next section) into a 75-min class. It would be possible to fit the activity into a 50-min class, however, by simply supplying students with an operational definition and by having them record data from one yearbook instead of two. I typically have students go through two because it provides more data and because it makes more of an impact to discover the same gender difference in two sources.

Reviewing Methodology

After students have had time to survey the data collected, I move on to a discussion of methodology. I state that, on the basis of the large gender discrepancy in smiling, we can reject our null hypothesis and accept our alternative. I then review why an operational definition is necessary. First, it makes it much easier to render the judgment about whether an individual is smiling. Second, it reduces potential bias: Without an operational definition, an ambiguous expression might be labeled smiling on a female student and not smiling on a male student.

Another methodological issue I discuss is interpretation of results. This point is especially important in gender research to overcome a tendency to explain gender differences based on a female deficit model (Hyde, 1996). For example, if men predict they will receive higher scores on a test, researchers have tended to conclude that women lack self-confidence rather than men have inflated egos. Similarly, some researchers (e.g., Deutsch, 1990; Henley, 1973, 1977) have explained gender communication differences in terms of women’s deficit in social status: Women smile more, gaze more, and interrupt less because of their inferior status relative to men. As an alternative to the female deficit model, other researchers have suggested that these same behaviors may occur for reasons such as women’s greater concern with interpersonal comfort and harmony or their different reactions to social unrest (Hall & Halberstadt, 1986; Tannen, 1991).

I ask students why they think female students smile more in yearbook photos. Are female students happier? Research on depression (Nolen-Hoeksema, 1990) would argue against this interpretation. Students are usually good at suggesting other interpretations, such as photographers encourage female students to smile more, female students like having their pictures taken more, and parents would not purchase photos of unsmiling daughters. Students also usually note that smiling is simply a more important part of the female role. For men, especially male high school students, smiling may even be viewed as inappropriate role behavior. Interestingly, research suggests that women’s smiling correlates less with their internal feelings than does men’s smiling (Bugental, Love, & Gianetto, 1971).

I also ask students if they can think of situations in which men would smile more often than women. Typical answers are “while watching a football game” or “while watching a violent or sexually explicit film.” This exchange leads to the issue of external validity: How generalizable are the results? The class data provide robust support for gender differences in smiling in yearbook photos (as differences are usually found across ages, types of schools, and time periods), but they do not demonstrate that women smile more in other situations. Thus, in terms of smiling in yearbook
photos, the external validity of the class’s research is high; in terms of smiling in general, the external validity is low. Although Hall (1984, 1987) found that adult women smiled more than men in 17 out of 18 studies, she acknowledged that her conclusions were limited by the fact that few gender and communication studies had been conducted in everyday, informal situations.

Other methodological issues can be addressed with this exercise. For example, instructors might want to have student pairs calculate intrarater reliabilities on their judgments, perform chi-square tests on their results, or calculate effect sizes. Students could also determine whether gender differences in smiling are larger in high school versus college yearbooks, in older versus recent yearbooks, or whether coder gender affects the results.

Evaluation

During intersessions in 1995 and 1996, 45 students (mostly juniors and seniors and about half psychology majors) enrolled in my Psychology of Gender course evaluated this exercise. They did this anonymously and voluntarily. First, students indicated how well the activity achieved its purpose, which I described as follows: “to give you ‘hands on’ experience investigating a gender difference and to highlight several methodological concerns (e.g., operational definitions, external validity, interpreting results).” Next, students indicated how much they enjoyed the activity. They answered both questions using 7-point scales. For the purpose item, 7 indicated extremely well and 1 indicated not at all well; for the enjoyment item, 7 indicated extremely enjoyable and 1 indicated not at all enjoyable. High ratings suggested that students thought the purpose was achieved (M = 6.02, SD = 1.03) and that the activity was enjoyable (M = 5.84, SD = 1.22). A third question asked “When I teach the course again, should I do this activity?” Of the 45 students, 43 indicated “yes.”

My original aim in asking these questions was to compare reactions to different activities in my gender course. Consequently, the questions were general and were not measures of actual learning. Instructors using this exercise in research methods might want to quiz students, before and after doing the activity, on their understanding of the methodological issues presented (null hypothesis, operational definition, external validity, etc.).

Instructors using this activity in any course should find that students regard the exercise as enjoyable. It provides a nice break from lectures and adds a bit of levity to the classroom. Students love seeing old photos of one another and observing individuals who look like famous pop stars or criminals. They also enjoy discerning the gender of androgynous students from the early 1970s and finding out how their instructor looked long ago.

References


Notes

2. I thank the editors and three anonymous reviewers for supportive comments and helpful suggestions.
3. Send correspondence to Angela Lipsitz, Department of Psychology, Northern Kentucky University, Highland Heights, KY 41099; e-mail: lipsitz@nku.edu.
I describe an in-class activity designed to meet the needs of research methods students who frequently encounter survey research and multiple regression. Groups of students receive the common task of developing a small number of questionnaire items to predict a specified criterion. Students then receive descriptive statistics and multiple regression results demonstrating the ability of their items to predict the specified outcome. In a sample of 8 groups, reactions to the activity were highly positive. In addition, students who participated in the activity (n = 30) scored slightly (although not statistically significant) higher on a subsequent exam than those in a control group (n = 18).

Despite their prevalence in published psychological research, simple regression, multiple regression, and survey design receive little attention in most introductory research methods texts. A cursory glance through several such texts (Bordens & Abbott, 1996; Graziano & Raulin, 1997; Leary, 1995; Ray, 1997) revealed an average of only 2.00 pages devoted to simple regression and 1.75 pages devoted to multiple regression. ANOVAs, on the other hand, received an average of 17.00 pages of coverage. Only two of the texts (i.e., Bordens & Abbott, 1996; Ray, 1997) devoted an entire chapter to survey design. The others included the subject with other “nonexperimental” procedures such as observation and case studies.

This emphasis on experimental design in textbooks often contradicts what students experience as they conduct literature reviews and routinely encounter articles that use regression and survey designs. The use of these methods is especially prevalent in traditionally popular areas such as drug abuse and child abuse where experimental research is rare or nonexistent. A search of PsycINFO (1887–1998) yielded 24 hits for “simple regression” and 4,801 hits for “multiple regression.” Using these search results as indicators of the prevalence of these procedures, it appears that students who rely on available texts would be ill prepared to understand correlational reports that use simple or multiple regression.

Teaching techniques related to regression and survey design are also conspicuously absent from Teaching of Psychology. A computer search of titles and abstracts since the journal’s inception revealed only two articles devoted to teaching regression. Although Kowalski (1995) recently described moderated multiple regression as an alternative to ANOVA in the analysis of mixed factorial designs, she cited no other research explicitly related to teaching regression. Williams (1975) also described a method for teaching students about multiple regression in the context of a statistics course. Survey design has only been explicitly addressed in one Teaching of Psychology article. Sattler, Back, and Pollitt (1995) reported that students who designed an exit survey benefitted from the experience through a better understanding of survey design, enhanced critical thinking, and a greater interest in research.

Based on my review of textbook coverage and available teaching techniques, there is a need for greater coverage of survey design and multiple regression in the research methods course. To address this issue, I developed an in-class activity that allows students to construct their own surveys and receive quick feedback regarding the overall predictive validity of the variables they chose. The primary goal of this exercise is to equip students to read and interpret the results of studies analyzed with multiple regression. Compared to Williams’s (1975) article, my focus is on the interpretation of the results rather than the calculation of the regression weights. Students should also gain an appreciation for the survey development process.

Method

Participants

Participants included 30 students (hereafter referred to as the activity group) enrolled in two different sections of a research methods course. Because of the scheduling of a holiday, a third section (n = 18) of the same course had one fewer class meeting, did not participate in the activity, and thus served as the control group.

Procedure

Participants in the activity and control groups received the same lectures on the basics of correlation, simple regression, multiple regression, and survey design. The lectures were loosely based on the corresponding sections of Research Methods (Bordens & Abbott, 1996). During the first class period after the lectures, I randomly assigned participants in the activity group to subgroups of three or four students. Each subgroup developed a questionnaire containing five items (or predictor variables) that would be administered to all students in the activity group. The goal was to generate five
items that (as a set in a multiple regression equation) would best predict a particular outcome. In this case the outcome (or criterion) to be predicted was individual performance (i.e., test scores) on the previous exam.

Students were free to generate any type of item that might predict an exam score. The only limitations were that (a) the items had to be either dichotomous or continuous (interval or ratio) in nature and (b) the item could not directly ask, “What was your score on the previous exam?” To increase motivation and interest in the activity, students learned that they would receive extra credit points on their next exam (up to 0.55% of total final grade) depending on the predictive validity of their regression equations.

To ensure that the control group had an equal chance at receiving extra credit points, I gave them a take-home quiz related to regression and survey design. In addition to questions regarding the appropriate use of open-ended and closed-ended survey items, the quiz required students to interpret and use multiple regression results.

After students in the activity group generated questionnaire items, I combined them into one questionnaire. I did not eliminate duplicate items so that a potential source of error (e.g., random responding) could be examined and demonstrated. I transcribed the items directly as the groups had constructed them. Thus students were responsible for generating survey items as well as response formats. During the next class period, students in the activity group completed the master questionnaire containing 40 items.

In the next class period, I gave each member of the activity group a detailed report that contained descriptive statistics for each item (i.e., means, standard deviations, frequency distributions), interitem correlations, and multiple regression results (i.e., b weights, beta weights, significance levels for each variable, Multiple R, R²). I used SPSS for Windows 6.1 (1994) to compute the results. The data from all groups were contained in the same file and coded by group to make data processing easier.

Students met in their original groups and discussed their results (i.e., the predictive ability of their five items). After discussion, each group elected a spokesperson who presented the group’s results to the class. The purpose of the presentation was to encourage students to make sure they fully understood their results. The questions they asked me while preparing for the presentation revealed occasional misunderstandings and gaps in their knowledge. During the presentation, students commented on their best predictors, the overall success of their models, and any surprises that they encountered. Groups received extra credit points based on the size of their Multiple R.

Measures

I evaluated the exercise by examining attitudes about the activity and performance on the next test. For the attitudes, students used a Likert scale to indicate their agreement or disagreement with statements such as “This activity should be used in future sections of this course” (see Table 1). The 7-point scales ranged from 1 (strongly disagree) to 7 (strongly agree). The test following the activity contained several multiple-choice items and several items requiring students to interpret and use the results of a study using multiple regression. Samples of the multiple-choice item stems were the following: “What is one drawback to using open-ended items?” “When using regression, what does the correlation between the actual Y and the predicted Y tell you?” In the other section of the test, students received a scenario and the results of a multiple regression analysis. Sample questions from this area were “Holding everything else constant, how much does Y change with each unit change in X2?” “Based on these results, what is the predicted Y for someone with predictor values of X1 = 10, X2 = 2.5, and X3 = 32?”

Results

Reactions to the activity were quite positive. The descriptive statistics presented in Table 1 reveal that students considered the activity worthwhile, believed that it increased their knowledge of both survey design and multiple regression, and thought that it should be used in future courses.

I also compared the test scores of students in the activity group with those of the control group. Although not statistically significant, t(46) = 1.02, p = .16, students in the activity group (n = 30, M = 79.79, SD = 15.01) scored somewhat higher on the portion of the exam dealing with survey design and regression than students in the control group (n = 18, M = 75.06, SD = 16.26).

Discussion

This activity served two primary functions. First, it demonstrated the challenges that survey researchers face. During item generation, many students remarked that it was difficult to agree on the five best predictors. Survey researchers likewise face the challenge of getting the most information out of the fewest number of questions (Sheatsley, 1983). Students also had to consider issues such as response format and respondent memory. During item generation, for example, several groups discussed variables that might have been valid predictors (e.g., hours of sleep the night before the exam) but subject to large amounts of error (Bradburn, 1983).

Second, the exercise likely gave students a better sense of “what the numbers mean” in multiple regression. As a specific example, the meaning of Multiple R is typically difficult for students to grasp. In this activity, however, students can...
readily see that there is some discrepancy between their actual exam score and the score predicted by their regression equation. They can also readily see that small discrepancies represent better prediction by their equation. Students take quite an interest in the Multiple $R$ because it is the criterion by which the different regression equations are compared.

Students also likely gained an appreciation for the ability of multiple regression to control for multiple variables at once. One group, for example, found a significant bivariate correlation, $r(34) = .48, p < .01$, between exam score and the item "It is very important that I get an A in this course." This group also included the significant predictor "This was an easy test," $r(34) = .44, p < .01$. When the variables were entered into the regression equation, however, perceived “easiness” of the test remained a significant predictor, but the importance of receiving an A was far from significant, $p > .42$. This example was particularly effective in showing the danger involved in trusting simple bivariate correlations.

Teachers who are more creative could devise additional criteria for prediction. I chose previous exam score because it was meaningful to the students and a wide variety of variables might predict it. Different criteria might increase the variety of predictors and make the exercise more interesting and challenging.

One drawback to this activity was the amount of time it required in and out of class. The item-generation period lasted approximately 45 min. Completing the survey during the next class period took approximately 10 min. Finally, I devoted an entire class period (i.e., 75 min) to the discussion of the results. In addition to the in-class time, entering the data, completing the analyses, and formatting the feedback handouts required approximately 4 hr of my time. Overall, I believe the time was well spent. The students certainly enjoyed it and a more powerful assessment of performance might show that they actually learned more than the control group.

Admittedly, the control group in this study was not a perfect one. It is possible that the activity group performed slightly better because of the extra time of exposure to regression. The quiz completed by the control group, however, provided some additional exposure. One possible explanation for the lack of a significant difference in the exam scores lies in the similarity between items on the control group’s take-home quiz and items on the exam. Specifically, students in the control group calculated a predicted $Y$ on both assessments. During the exercise, students in the activity group never calculated a predicted outcome (i.e., grade). Because these exam items were embedded and scored within more comprehensive questions, it is impossible to determine the effect of this confound. Nevertheless, the results indicate that students found the activity useful and also benefitted from it. I hope that other teachers also will find it to be a useful tool for teaching two traditionally neglected topics.

**References**


**Note**

Send correspondence to Thomas A. Timmerman, Department of Psychology, Austin Peay State University, 601 College Street, Clarksville, TN 37044; e-mail: timmermant@apsu.edu.
In this article, I describe how I used a portfolio assignment to facilitate learning and provide a useful resource for future research experiences. Forty-eight students from 2 introductory methods and statistics courses compiled a portfolio that reflected their individual approach to understanding the major topics of the class. The portfolio assignment required students to integrate material from the text, class lectures, and laboratory assignments into a resource that was both personally relevant and accessible. Student feedback, performance data, and thesis advisor ratings supported the utility of this assignment.

A variety of methods of enhancing learning in statistics and research methods courses have been documented (Beins, 1985; Derry, Levin, & Schauble, 1995; Dillbeck, 1983; Thompson, 1994). However, many of these methods do not necessarily extend beyond the introductory methods and statistics courses. Conners, Mccown, and Roskos-Ewoldsen (1998) suggested that instructors need to intentionally integrate statistics throughout the curriculum. Such integration may allow for more meaningful research opportunities like the baccalaureate thesis (Terry, 1983). In this article, I describe a portfolio assignment that I designed to facilitate learning and encourage continuity between an introductory methods and statistics course and subsequent research experiences.

Throughout an introductory research methods and statistics class, students compiled a portfolio that reflected their individual approach to understanding the major topics of the class. I described the rationale and procedure for the portfolio assignment in the course syllabus. The ultimate goal was for students to develop a personalized resource that was useful not only for this specific course, but also for future research opportunities. More specifically, the portfolio serves two primary functions: (a) to enhance learning by requiring students to actively integrate the class material into a concise resource and (b) to give students a personalized and accessible tool to use in future research endeavors.

At Westminster College, all psychology majors must complete a senior thesis in which they design, conduct, analyze, and present a novel research project. By the time they reach the senior thesis, many students find it difficult to retrieve information learned in their introductory methods and statistics course. Students ultimately return to their notes and texts to help clear the cobwebs. The portfolio assignment attempts to bridge the gap between the introductory methods and statistics course and the senior thesis by providing a more individualized and thus more understandable resource.

Method

Participants

I used the portfolio assignment in two sections of a one semester integrated research methods and statistics course. This course serves as a prerequisite for all advanced laboratory courses as well as the senior thesis and is required of all psychology majors and minors. There were 25 students (10 men, 15 women) in the first section (Fall, 1997) and 23 students (6 men, 17 women) in the second section (Fall, 1998). Each of the students in the two classes was either a psychology major (71%) or a psychology minor (29%).

Procedure

The procedure for the portfolio assignment was identical for the two sections. However, I provided the second section (Fall, 1998) with comments from the previous year's students to give them ideas for their portfolio. At the beginning of the semester, I instructed the students that a useful portfolio would integrate material from the text, class lectures, and laboratory assignments in a way that was both personally relevant and accessible. To promote active integration and consideration of the material to include in the portfolio, I imposed several guidelines on portfolio construction. First, I provided a list of required and suggested topics in the syllabus (see Table 1). To prevent the “shotgun” approach in which students simply include as much information as possible, I also placed a page restriction on each topic contained in the portfolio. Specifically, students could include only three pages of material for each of the topics specified in the syllabus. I also encouraged students to avoid a literal translation of the text or the notes and to make their portfolio unique by using many examples, particularly ones that were relevant to their interests.

Aside from these recommendations and requirements, I gave students considerable flexibility in the construction of their portfolio. I checked each student’s progress on the portfolio several times during the semester, but I did not grade these periodic checks. I simply provided ideas and feedback on the accuracy and relevance of the material they had chosen to include.
Evaluation and Conclusions

Portfolio Quality

To evaluate the quality of student portfolios, an independent rater (blind to the student’s course performance) reviewed the content of the portfolios. The rater was a colleague who teaches Westminster’s senior thesis seminar course as well as a different section of the introductory methods and statistics class. The rater evaluated portfolios only from the second section (Fall, 1998). Therefore, all analyses involving portfolio quality ratings involve only that second section. All other analyses, unless otherwise stated, involve both sections.

The rater evaluated the portfolios along three dimensions: accessibility, anticipated value, and uniqueness. Accessibility refers to the ease with which the rater could find specific information that might be needed in a research study (e.g., which analysis to use, how to interpret printouts). Anticipated value refers to the rater’s assessment of how useful the portfolio would be to the student during the senior thesis project. Uniqueness refers to the extent to which the student went beyond the materials provided by the instructor and the textbook (e.g., new examples, annotations, added emphasis). Because the uniqueness of the portfolio content would not be easily discernable to an outside reader, I (instructor) supplied the ratings on this dimension prior to grading the final exams.

Performance Data

One of the primary goals of the portfolio was to facilitate learning of the course material. To examine the relation to learning, I looked at portfolio quality ratings and student performance on a cumulative final and three in-class exams.

<table>
<thead>
<tr>
<th>Table 1. Required and Suggested Topics for the Portfolio Assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Section</td>
</tr>
<tr>
<td>--------------------------------</td>
</tr>
<tr>
<td>Descriptive analyses</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Inferential analyses</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Experimental design</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Optional topics</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Note. Each of the topics can be a maximum of three pages.

Cumulative final exam. At the end of the semester, students took a cumulative final exam using only their portfolio as a resource. The final exam lasted approximately 2 hr and contained problems that were primarily applied and integrative in nature. Specifically, I gave students descriptions of experimental studies and excerpts from computer printouts and then asked various questions about the design, statistical analysis, and interpretation of the results. I used the final exam as a criterion because it not only reflected the content of the class but it also accurately represented the skills needed by students during their senior thesis (e.g., choosing analyses, identifying confounds, interpreting printouts, reporting results). Theoretically, students who prepared more accessible and useful portfolios would perform better on the research skills targeted by the final exam.

Final exam performance was positively correlated with ratings of the portfolio’s accessibility, $r(21) = .50, p = .015$; anticipated value, $r(21) = .56, p = .005$; and uniqueness, $r(21) = .65, p = .001$, suggesting that students who prepared higher quality portfolios performed better on the research skills measured by the final exam. Because performance on the final is also likely to be affected by general ability and evaluation anxiety, I collected each student’s cumulative grade point average (GPA) as well as his or her score on the evaluation anxiety scale of the Math Anxiety Rating Scale–Revised (Plake & Parker, 1982). Controlling for GPA and math evaluation anxiety, final exam performance was still positively and moderately related to the portfolio’s accessibility, $r(18) = .55, p = .013$; anticipated value, $r(18) = .54, p = .015$; and uniqueness, $r(18) = .67, p = .001$. Thus, it appears that a well-constructed portfolio may be a valuable reference for the types of tasks represented in the final exam.

In-class exams. Because I told students they would be able to use only their portfolio for the final exam, the relation between portfolio quality and final exam performance is not surprising. However, performance on the three in-class exams served as a measure of learning that was independent of portfolio use. Although students compiled their portfolios throughout the semester, I did not permit them to use any outside resources (including the portfolio) for the in-class exams. After controlling for GPA and math evaluation anxiety, performance on the three in-class exams was positively and moderately related to the accessibility, anticipated value, and uniqueness of the portfolio (see Table 1).

These data suggest that the process of organizing, evaluating, and selecting material for the portfolio may facilitate students’ understanding of the material. However, these data are correlational in nature, and experimental studies are needed to evaluate any causal effects of the portfolio on learning.

Follow-Up Data

Data from the most recent senior thesis class (Fall, 1999) suggested that the portfolio was a useful research tool beyond the introductory class. Specifically, as a part of a departmental assessment of the senior capstone course, each thesis advisor rated the amount of methodological and statistical assistance required by each of his or her advisees on a scale ranging from 1 (no assistance at all) to 7 (extensive assistance). Advisor ratings
indicated that students who completed the portfolio assignment \( (M = 2.22, SD = 1.09, N = 9) \) required significantly less assistance than those who did not have a portfolio \( (M = 3.50, SD = 1.20, N = 8) \), \( t(15) = 2.30, p = .036 \).

I also solicited anonymous comments from students after they completed their senior thesis. In general, these students reported that the portfolio was very useful during their senior thesis. Some students stated that the portfolio was helpful for parts of their senior thesis but not for others (i.e., more advanced designs and analyses). Other students stated that they would approach the portfolio differently if they were to do it again. For example, several students acknowledged that they mistakenly left some information out (step-by-step procedures for statistical programs, reading printouts) because they thought they would remember it later. Each year, I share these comments as well as sample portfolios with the next class.

### Student Evaluations

At the end of each semester, students rated the educational value of the portfolio assignment on a scale ranging from 1 (not at all valuable) to 7 (extremely valuable) and made recommendations for future use. For the first section (Fall, 1997), the mean rating of educational value of the portfolio assignment was 5.06 \( (SD = 1.58, N = 25) \). Additionally, approximately 74% of the class recommended using the portfolio assignment again either "as is" (65%) or "with modifications" (9%). The mean rating of educational value for the second section (Fall, 1998) was 5.77 \( (SD = 1.27, N = 23) \). Unlike the previous semester, each of the students in the class recommended using the portfolio assignment again either "as is" (91%) or "with modifications" (9%).

Psychology majors \( (M = 5.53, SD = 1.59) \) and nonmajors \( (M = 5.12, SD = 1.19) \) did not differ significantly on their ratings of the educational value of the portfolio, \( t(43) = .85, p = .40 \). However, in the end-of-semester evaluations each year, several students requested that psychology minors be exempt from the portfolio requirement or that the workload be diminished for minors. One reason for these comments may be that psychology minors at Westminster are not required to complete a senior thesis. Psychology majors, on the other hand, may have believed that the amount of effort invested in the portfolio was justified given the impending senior thesis.

Qualitative data (end-of-semester anonymous comments) suggested that students perceived the portfolio as time intensive but valuable. Student comments also emphasized the difficult nature of deciding what to include and what not to include in the portfolio. Although making choices about what to include may be frustrating for the student, it is a dilemma that most instructors would welcome because it requires the student to evaluate the material rather than simply include it.

From an instructor’s perspective, the most problematic aspect of the portfolio assignment was getting students to move beyond the specific material and examples covered in class or in the laboratory. Some students found it difficult to generate new examples or supplement class materials with personal annotations. I found that this reluctance to move beyond class examples or notes was generally a good indication that the student had not mastered that material and needed additional assistance. Thus, the portfolio became a valuable formative assessment tool for me. In addition, the portfolio assignment is potentially dynamic and flexible. As students progress through the psychology curriculum, they can add information from upper level courses and use the portfolio for research-oriented class projects.

In sum, both quantitative and qualitative data suggest that the portfolio may be a useful way of facilitating learning and providing connections to future research experiences. Because the portfolio assignment is time intensive, it may not be desirable for all students (viz., nonpsychology majors). However, data from this study suggest that the value may be substantial for students who go on to complete a baccalaureate thesis or other independent research.

### Notes

1. Mark J. Sciutto is now at Muhlenberg College.
2. Portions of this article were presented at the annual meeting of the American Psychological Association, Boston, August 1999.
3. I thank Sandra K. Webster for her assistance in evaluating the portfolios and for her comments on an earlier version of this article. I also thank the editor and three anonymous reviewers for their helpful suggestions on earlier drafts of this article.
4. Send correspondence to Mark J. Sciutto, Department of Psychology, Muhlenberg College, Allentown, PA 18104; e-mail: sciutto@muhlenberg.edu.

### References


### Table 2. Partial Correlations Between Portfolio Quality Ratings and In-Class Exam Performance (Controlling for Grade Point Average and Math Evaluation Anxiety)

<table>
<thead>
<tr>
<th>Portfolio Dimension</th>
<th>Exam 1</th>
<th>Exam 2</th>
<th>Exam 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accessibility</td>
<td>.55*</td>
<td>.82***</td>
<td>.68**</td>
</tr>
<tr>
<td>Anticipated value</td>
<td>.62**</td>
<td>.76***</td>
<td>.58**</td>
</tr>
<tr>
<td>Uniqueness</td>
<td>.30</td>
<td>.51*</td>
<td>.39</td>
</tr>
</tbody>
</table>

**Note.** Students did not use the portfolio for the three in-class exams. *p < .05. **p < .01. ***p < .001.
Learning By Doing: Research Methods With a Theme

Pam Marek  
Anderson College

Andrew N. Christopher  
Albion College

Bonnie J. Walker  
University of Central Florida

Embracing an active learning approach, a 1-semester research methods course with a statistics laboratory component described herein provided sophomore students with their initial and only required exposure to research methods. Techniques used overcome constraints of limited library resources, restricted interlibrary loan procedures, and lack of a formal participant pool. The instructor selected the theme for a single class project. Students participated in design, data collection, analyses, and report writing. Discussion focuses on the advantages and disadvantages of involving all students in the same project.

To provide hands-on methodological experience and application of statistical concepts, instructors have engaged students in individual projects (Smith, 2002); yoked statistics and research design courses together, with students working in dyads on student-chosen topics (Bhalla & Cornell, 2002); and involved an entire class in a single community-based research project (Chapdelaine & Chapman, 1999). The Smith course was third in a sequence of four research-oriented courses all taught in the psychology department. The Bhalla and Cornell innovation, targeted at third-year students with three psychology course prerequisites, involved two instructors with students receiving credit for two courses. The Chapdelaine and Chapman course involved two instructors, teaching and laboratory assistants, and personnel from a community agency to support and supervise the research.

The course described herein, taught by a single instructor, adapted this active learning approach to a small 4-year college that requires only a single sophomore-level research methods course with a 2-hr per week laboratory component, preceded by a generic statistics experience taught by mathematics faculty. Constrained by limited library resources, restricted interlibrary loan access, and lack of a formal participant pool for psychological research, all students participated in a single class project. However, given the lack of a teaching assistant or laboratory instructor that Chapdelaine and Chapman (1999) deemed essential for their community-based research, students conducted campus-based research revolving around a single theme.

Using a single theme provided considerable continuity to the course. The first and second authors have each incorporated a theme and related background readings into several class discussions of research design, American Psychological Association (APA)-style writing activities, and laboratory sessions. The first author has taught a theme-based course with a single class project four times with three different themes (the academic ethic [used twice]; religious orientation and materialistic values; and procrastination, the Protestant work ethic, and personality). The second author has taught such a course once (with a theme relating to affluence cues and person perception). Instructors chose themes based on personal interests and student suggestions from prior semesters.

The first author initially implemented the academic ethic theme at another 4-year college, in the first semester of a two-semester integrated statistics–methods sequence (the only required research-focused courses) that did not include a formal laboratory section. At both schools, students conducted a survey with dormitory residents as participants, individually drafted report sections, analyzed data using SPSS, and worked in dyads to write final reports. The subsequent course description depicts the one-semester methods course with a laboratory component.

Precourse Preparation

Given limited library facilities, lead times, and costs for interlibrary loans, the instructor organized reading materials before the course began, placing copies on library reserve. Given time constraints and other required topic coverage during the course, the instructor also developed a relatively complete draft questionnaire, including three preexisting scales, suitable for administration to students in the course. Finally, the instructor prepared theme-oriented lectures that extracted basic information from prior studies exploring theme-related issues and involving scale development.

Course Objectives Related to Theme

Course objectives reflected specifically in the class research project targeted several of the skills (e.g., research methodology and statistics, technology and computer, written and oral communication, critical thinking and problem solving) included in the Kruger and Zechmeister (2001) skill inventory for undergraduate psychology majors. Theme-related assignments aimed to enable students to apply concepts of design, implementation, analysis, and interpretation to psychological research; understand the role of statistics in scientific analysis; facilitate computer-based calculation of statistics; communicate research findings; and critically evaluate the design and interpretation of research. Evaluative items assessed student judgments of whether they reached these targets.

Implementing the Theme

During the first week, all students completed the draft questionnaire, familiarizing them with content, stimulating discussion, and providing pretest data for initial laboratory activities. Working backward from the questionnaire, students learned to develop research questions. For background knowledge, they read a comprehensive introduction from a relevant article (Rau & Durand, 2000), concerning academic behaviors, attitudes, and grades. Students discussed how the questionnaire applied to these behaviors and created hypotheses to test in a survey of resident students. Based on their hypotheses, students critiqued the questionnaire they pretested. The instruc-
tor then presented lecture material highlighting basic
information about prior research and the scales. Students
raised questions about specific scale items, eliminated some
behavioral questions, and added others. Thus, students gained
at least a partial feeling of “ownership” in the questionnaire.

During coverage of research ethics, students prepared an
informed consent form, discussing the elements of informed
consent and completing a form suitable for submission to an
Institutional Review Board. The instructor assumed respon-
sibility for ensuring compliance with ethical guidelines. The
instructor asked each student to collect data from five re-

donents. To develop a sampling scheme, student volunteers
obtained a floor plan indicating the number of residents
on each dorm floor. Students determined how many respon-
dents from each dorm should be asked to participate and how
to draw potential respondents as randomly as possible from
each dorm floor. Randomly assigned to dorm floors, students
spent 3 weeks collecting data.

Prior to analysis, students studied potential pitfalls in anal-
ysis of correlational research, third variables, bidirectionality,
outliers, and statistical control techniques. Given the volume
of data and other components of student workload, when
students returned questionnaires, the instructor entered the
data. Students discussed issues relating to illegibility, missing
data, or haphazard responding during class. Whereas stu-
dents practiced conducting and interpreting computer-based
analyses during laboratory sessions, the instructor provided
all students with the same comprehensive packet of output.

Laboratory Component

Students convened in the computer laboratory once a
week for 2 hr, primarily to learn how to conduct and interpret
statistical analyses using SPSS. As shown in Table 1, labs en-
compassed a relatively standard set of statistical procedures.
The instructor devised topic sequencing to ensure that stu-
dents were familiar with the statistical tests needed to analyze
data from the class survey. Because relevant analyses were
correlational, labs emphasizing ANOVA did not incorporate
class project data.

### Written Assignments Related to the Theme

Students learned about the various sections of an
APA-style report in a step-by-step manner. Using primary
source articles, several of which were related to the class pro-
ject, students completed a series of exercises that generally
cluded reading and identifying specific components of a ti-
tle page, abstract, introduction, method, results, discussion,
and references section. For example, the methods exercise
cluded an article with several scales in which students iden-
tified required information about each (e.g., what it mea-
ures, source, number of items, type of scale, range of scale
and anchor labels, sample item, reliability and validity infor-
mation). Additionally, for the results assignment, students
read the results section of the article that served as the inspira-
tion for the class project (Rau & Durand, 2000), answered
questions about statistical techniques and figures, and identified
formats for basic statistics.

After completing the relevant exercise, students worked in-
dividually to prepare drafts of each section of the class project
report, selecting articles from folders on library reserve. In the
report, the instructor required students to cite several key arti-
cles (e.g., sources of the scales) plus other articles that students
chose from a menu of topics (e.g., need for cognition, locus of
control). The instructor sequenced the draft assignments as
follows: title page and references, method, introduction (due
the ninth week of the semester), results, discussion, and ab-
stract. The instructor led class discussions of problems that

<table>
<thead>
<tr>
<th>Session</th>
<th>Topic</th>
<th>Link to Theme</th>
<th>Homework Assignmenta</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction to SPSS and data entry</td>
<td>Pretest data used in class</td>
<td>Computing frequencies, descriptive statistics, making histograms, and answering questions about output</td>
</tr>
<tr>
<td>2</td>
<td>Frequencies, descriptive statistics, histograms, selecting cases</td>
<td>Theme-relevant data used for homework</td>
<td>Writing results section to report descriptive statistics for scales and correlations</td>
</tr>
<tr>
<td>3</td>
<td>Correlation and partial correlation</td>
<td>Theme-relevant data used for homework</td>
<td>Writing results section to report chi-square test and preparing table</td>
</tr>
<tr>
<td>4</td>
<td>Chi-square</td>
<td>Theme-relevant data used in class</td>
<td>Writing two results sections to report independent and dependent t tests</td>
</tr>
<tr>
<td>5</td>
<td>t tests, independent and dependent</td>
<td>Final data from class project used in class</td>
<td>Writing materials subsection of method section to report information about scales</td>
</tr>
<tr>
<td>6</td>
<td>Working with scales (reverse scoring, summing scale scores, item–total correlation, split-half reliability, Cronbach's alpha)</td>
<td>Final data from class project used for homework</td>
<td>Writing results section to report data relating to prediction equation for college GPA</td>
</tr>
<tr>
<td>7</td>
<td>Simple regression</td>
<td>Final data from class project</td>
<td>Writing results section for each ANOVA lab to report results of omnibus and post hoc tests</td>
</tr>
<tr>
<td>8</td>
<td>One-way ANOVA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Factorial (two-way) ANOVA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Repeated measures ANOVA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

All assignments required computer-based data analysis and answering questions about computer output. The written assignments for Lab 3 through Lab 10 also required a minimum of two paragraphs written in APA style, including appropriate statistics.
plagued multiple drafts and offered considerable individual feedback on draft assignments. After students had submitted individual drafts of virtually all sections, the students worked in dyads to complete final reports, due the last day of class.

**Course Evaluation**

To evaluate the extent to which the theme-based class project supported course objectives, students used a 7-point scale to assess course effectiveness on dimensions specifically linked to the class project. Table 2 reveals that students rated the concept of a class project with an academic ethic theme favorably, recommended including it in future courses, and perceived that objectives were fulfilled. Ratings for a subsequent course with a theme involving religious orientations and materialistic values generally paralleled those for the course with the academic ethic theme. Interpretation of these data should be tempered by the consideration that students might routinely believe they learned a lot in a course. Furthermore, because comparable data for a similar course without a theme were not available, it is difficult to assess the extent to which the class project per se contributed to these ratings.

More subjectively, the quality of final reports authored by students previously unfamiliar with psychological research and APA style also suggested course effectiveness. Moreover, a few students subsequently elected to engage in independent research, a very seldom-chosen option prior to the introduction of this course format. Such student initiative may hint that a hands-on research methods course with a theme nurtured understanding of and comfort with the research process. Indeed, future research might address the extent to which skills taught in a theme-based research methods course have transferred to subsequent courses.

**Conclusions and Caveats**

A course including a class project with a theme is well suited to a one-semester, sophomore-level course, with a single instructor at a small school with limited library resources and no participant pool, providing a facsimile of the research process and a scaffold for subsequent individual projects. By incorporating a single theme, the instructor overcame potential difficulties arising from students obtaining and reading sufficient primary source material, developing a research idea, preparing materials, and conducting an individual project within the time frame of a single-semester course. Engaging students in a single project also facilitated obtaining sufficient participants for meaningful analyses when a participant pool is not available. A preselected set of primary source articles also enhanced the instructors’ ability to detect unintentional plagiarism and to address techniques for integrating material and expressing ideas in students’ own words. Moreover, instructors can assist the class as a group in understanding the complexities of primary source material and relating it to standard methodology and statistical topics. Students who might otherwise be detached from what they perceive as “dry” lecture material tend to be more receptive to learning concepts that are linked to research they are actually doing (e.g., Goodwin, 2002). Finally, frequent classroom discussion of topical issues linked to the theme and detailed critiques of draft sections provide even the weaker students with adequate conceptual understanding to prepare an appropriate final report. Whereas we acknowledge that students in research methods courses in small colleges successfully complete individual projects on varying topics, a single theme may help engender a climate geared to success.

However, a disadvantage of the single theme approach with articles on reserve is that students do not necessarily have an opportunity to do a literature search and develop their own research ideas. Moreover, given the detailed instructor feedback, students may not be considered sole authors of their final manuscripts. Additionally, successfully implementing “learning by doing” in the context of teaching standard methods material required considerable time and patience at several points. Furthermore, the workload prompted student complaints at the beginning of the semester, before students realized how much they would learn.

**Table 2. Specific Ratings of Class Project**

<table>
<thead>
<tr>
<th>Evaluative Item</th>
<th>Fall 2001</th>
<th></th>
<th>Fall 2002</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Overall, I would recommend including a class project in future sections of this course.</td>
<td>5.74</td>
<td>0.73</td>
<td>5.71</td>
<td>1.20</td>
</tr>
<tr>
<td>Objective 1: Apply concepts of design, implementation, analysis and interpretation to psychological research</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The class project helped me understand how research is done in psychology.</td>
<td>6.05</td>
<td>0.91</td>
<td>5.71</td>
<td>1.07</td>
</tr>
<tr>
<td>I like the idea of actually doing a research project rather than just reading about how research was done.</td>
<td>5.53</td>
<td>1.07</td>
<td>5.21</td>
<td>1.93</td>
</tr>
<tr>
<td>Objective 2: Enhance understanding of role of statistics in scientific analysis.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The class project helped me understand the role of statistics in scientific analysis.</td>
<td>6.11</td>
<td>0.74</td>
<td>5.29</td>
<td>1.73</td>
</tr>
<tr>
<td>Objective 3: Facilitate computer-based calculation of statistics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>By the end of the course, I was more comfortable doing computer-based analyses than I was at the beginning.</td>
<td>5.95</td>
<td>0.97</td>
<td>6.07</td>
<td>0.92</td>
</tr>
<tr>
<td>Objective 4: Strengthen written and oral skills required to communicate research findings</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The class project enhanced my ability to write in APA style.</td>
<td>6.28</td>
<td>0.81</td>
<td>6.00</td>
<td>1.18</td>
</tr>
<tr>
<td>Putting the final report together gave me a feeling of accomplishment.</td>
<td>6.26</td>
<td>1.04</td>
<td>6.50</td>
<td>0.76</td>
</tr>
<tr>
<td>Objective 5: Develop critical thinking skills relevant to design and evaluation of research.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>To understand the class project, I had to think about things I hadn’t thought about before.</td>
<td>6.56</td>
<td>0.69</td>
<td>6.29</td>
<td>0.99</td>
</tr>
</tbody>
</table>

*Note.* Ratings were based on a scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*).

Finally, because library regulations permitted use of instructor-provided reserve material for one semester only, successive course iterations at the same institution require either a new theme or the collection of permissions from publishers to assemble material into course packets. These difficulties, however, may be viewed as problems that are resolvable. An empirically testable question remains concerning the long-term effectiveness of learning in classes with a single theme-based project versus classes with individual projects. However, we contend that the challenge of fostering a climate of learning by doing is well justified by the pride students display in their final reports, the quality of the written documents, and the potential for voluntary student involvement in future research.

References


Notes

1. Portions of this article were presented at the 24th National Institute on the Teaching of Psychology, St. Petersburg Beach, FL, January 2002.
2. We thank Randolph Smith and three anonymous reviewers for their comments on an earlier version of this article.
3. Send correspondence to Pam Marek, Department of Psychology, Anderson College, 316 Boulevard, Box 1166, Anderson, SC 29621, e-mail: pmarek@ac.edu.

PSI and Distance Learning in a Developmental Psychology Course

Marion A. Eppler and Marsha Ironsmith

East Carolina University

This article describes a Web-based mastery course in human development. We describe the components of our Personalized System of Instruction (PSI) course, compare the effectiveness of this format to our lecture sections, and conclude with a discussion of the goodness of fit between PSI and distance learning. Consistent with previous research, Web PSI students exceeded lecture students on almost every outcome measure. We argue that increasing demand for distance learning sets the stage for a revival of PSI and that a combination of PSI and Web formats fits the needs of a growing population of continuing education and nontraditional college students.

Keller (1968) introduced the concept of a Personalized System of Instruction (PSI) in his classic article, “Good-bye, Teacher ....” The features that distinguish PSI from traditional instruction include emphasis on written materials rather than lecture, small units with frequent quizzing, immediate feedback, mastery criterion, self-pacing, and undergraduate proctors (Keller, 1968; Kulik, Jaksa, & Kulik, 1978). The impressive body of research on PSI documents its efficacy as a pedagogical tool (Kulik et al., 1978; Kulik, Kulik, & Cohen, 1979; Robin, 1976; Sherman, 1992). Despite clear evidence of success, use of PSI in the classroom and research on PSI have declined sharply since the early 1980s (Buskist, Cush, & DeGrandpre, 1991; Novak, 2001). Buskist et al. blamed inertia of the educational system for the decline in popularity of PSI.

College students today differ considerably from students in the 1960s and 1970s when PSI flourished. Fewer fit the traditional pattern of 18- to 22-year-olds who attend full time and live on campus. Universities now serve more nontraditional students who work, have family obligations, and want greater flexibility in class scheduling (Levine & Cureton, 1998). Some universities have promoted Internet-based instruction as a way of catering to the needs of nontraditional students. PSI could be a valuable means of delivering information online, thus integrating an old and proven pedagogy with modern technology.

We designed a developmental psychology course combining PSI techniques with Internet resources. To evaluate the effectiveness of our Web PSI course, we compared it to our conventional lecture-based instruction.

Method

Participants

Participants were 814 students enrolled in 14 sections of our course across four semesters. As seen in Table 1, Web PSI and lecture students were comparable in terms of self-reported cumulative grade point averages (t < 1.0, for each of the four semesters).

Procedure

We each taught a Web PSI section all four semesters (eight sections total), and we conducted these sections collaboratively. We each taught a lecture section for the first two semesters and alternated for the next two semesters (six sections total). All students used the same textbook and took the same comprehensive final examination.
METHODS AND TECHNIQUES

Mythbusters: A Tool for Teaching Research Methods in Psychology

Edward Burkley and Melissa Burkley
Oklahoma State University

Mythbusters uses multiple research methods to test interesting topics, offering research methods students an entertaining review of course material. To test the effectiveness of Mythbusters clips in a psychology research methods course, we systematically selected and showed 4 clips. Students answered questions about the clips, offered their opinions of them, and ultimately responded to Mythbusters-related exam items. Students indicated that the clips helped them understand course concepts, helped them apply their knowledge to real-world research studies, and provided them with an enjoyable educational experience. More important, students performed better on Mythbusters-related exam items than on control items, suggesting the clips were effective pedagogical tools for psychology research methods courses. We discuss specific Mythbusters clips that are relevant to psychological concepts.

I do all my own experiments.
—Mythbusters

The use of popular media, from comics to films, can be an enjoyable and effective medium for conveying class material to students. Across the field of psychology, numerous articles have attested to the teaching benefits of using television clips and movies in the classroom (e.g., Anderson, 1992; Bolt, 1976; Boyatzis, 1994; Christopher, Walter, Marek, & Koenig, 2004; Conner, 1996; Dorris & Ducey, 1978; Eaton & Uskul, 2004; Fleming, Piedmont, & Hiam, 1990; Roskos-Ewoldsen & Roskos-Ewoldsen, 2001). One reason for this abundance of articles might be that, unlike other disciplines, psychological concepts are readily accessible through film. However, research methods courses in psychology often fail to capitalize on the benefits of film clips.

Research methods courses typically cover what students consider dry material, and as a result, instructors are often looking for ways to increase their students’ interest. For example, instructors have used handwriting analysis (Boyce & Geller, 2002), pseudoscientific claims (LoSchiavo & Roberts, 2005), and online lab chat rooms (Sommer & Sommer, 2003) in an attempt to make research methods courses more accessible, meaningful, and applicable to students. In this article, we propose that clips from the Discovery Channel’s show Mythbusters are another tool that instructors can use to increase students’ interest in and knowledge of research methods.

Mythbusters is a television show that explores various myths, from ancient legends to modern folklore, through empirical means. In each episode, hosts Adam Savage and Jamie Hyneman test myths using a variety of research methods to determine if the myth is “confirmed” or “busted.” The Mythbusters’ use of various research methods makes the film clips quite compatible with the material in a psychology research methods class.

We presented several clips from the Mythbusters show in our undergraduate research methods course as a means for students to better learn the course material. To select the clips, we limited our choices to episodes available on the Mythbusters: Collection 1 DVD (Rees, Luscombe, Rudolph, LeDonne, & Plavnick, 2007) because this was the only Mythbusters DVD collection available at the time. We then identified clips (or
myths) for inclusion based on three criteria. First, the clips had to highlight key concepts that are standard in any psychology research methods course (e.g., measurement, experiments, etc.). Second, the clips had to vary in the type of design or method used to test the myth (e.g., factorial design, within-subjects design, etc.). Finally, the clips had to be enjoyable to students, so we chose clips that tested the more interesting myths. We selected four clips that fulfilled these criteria (see Table 1).

We showed the four *Mythbusters* clips throughout the semester, one clip every 2 to 3 weeks. The first clip tested the myth that when buttered toast falls on the floor, it always lands buttered-side down. For their study, the researchers designed several toast-dropping machines and completed several trials with buttered toast and nonbuttered toast. The clip provided examples of how to calibrate measurement instruments, the benefits of a control group, and ways to reduce measurement error. Students viewed the clip after our lecture on psychological measurement.

The second clip tested the myth that an individual is more likely to yawn after seeing someone else yawning. Rather than using the whole clip, we focused on the demonstration used in the last half of the clip because it illustrated a better research design. To test this myth, an experimenter greeted participants and either did or did not yawn while explaining the study. Participants were then seated in isolated rooms to wait for further instructions. Hidden cameras were used to see if the participants subsequently yawned while they waited for the experimenter to return. The clip highlighted the use of control and experimental conditions and provided an illustration of an observational study with unobtrusive behavioral measures. Students viewed the clip after our lecture on observational methods.

The third clip tested the myth that talking on a cell phone while driving as bad as driving drunk. To test this myth, two participants were tested on a driving course under three different conditions: control, driving while talking on a cell phone, and driving while drunk. The clip provided an illustration of a within-subjects experimental design and highlighted issues related to this design. Students viewed the clip after our lecture on the distinction between within-subjects and between-groups experimental designs.

The final clip tested the myth of who would be wetter—a person who runs in the rain or a person who walks in the rain. To test this myth, the researchers created an elaborate rain-making machine in an airplane hanger to control for various rain and wind characteristics. The participants then either walked or ran through the rain while a fan was or was not blowing (to simulate wind). The researchers wore white coveralls (over latex suits to control for perspiration) to assess the amount of rain absorbed. After each trial, the researchers weighed the coveralls to determine how wet they had become. This clip provided an example of a $2 \times 2$ factorial design and clearly demonstrated the concepts of main effects

### Table 1. *Mythbusters* Clips Used From Seasons 1 Through 3

<table>
<thead>
<tr>
<th>Episode No. and Clip Title</th>
<th>Research Methods Concepts</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>28: Does toast always land buttered-side down?</td>
<td>Nominal scale of measurement&lt;br&gt;Measurement error&lt;br&gt;Instrument calibration&lt;br&gt;Control group</td>
<td>11 min</td>
</tr>
<tr>
<td>28: Is yawning contagious?</td>
<td>Independent and dependent variables&lt;br&gt;Between-groups experimental design&lt;br&gt;Operational definitions&lt;br&gt;Control group&lt;br&gt;Observational research measure</td>
<td>4.5 min</td>
</tr>
<tr>
<td>33: Is talking on a cell phone while driving as bad as driving drunk?</td>
<td>Independent and dependent variables&lt;br&gt;Within-subjects experimental design&lt;br&gt;Carryover effects&lt;br&gt;Counterbalancing&lt;br&gt;Experimenter expectations</td>
<td>17 min</td>
</tr>
<tr>
<td>1: Who gets wetter? (running vs. walking in the rain)</td>
<td>Independent and dependent variables&lt;br&gt;Factorial ANOVA design&lt;br&gt;Main effects and interactions&lt;br&gt;Internal and external validity&lt;br&gt;Control over extraneous variables</td>
<td>13.5 min</td>
</tr>
</tbody>
</table>
Table 2. Student Evaluations of Mythbusters Clips

<table>
<thead>
<tr>
<th>Item</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The Mythbusters clips helped me to understand the class material better.</td>
<td>6.1</td>
<td>0.69</td>
</tr>
<tr>
<td>2. The Mythbusters clips were a good supplement to lecture.</td>
<td>6.4</td>
<td>0.67</td>
</tr>
<tr>
<td>3. The Mythbusters clips made the concepts discussed in class seem more &quot;real-world.&quot;</td>
<td>6.4</td>
<td>0.74</td>
</tr>
<tr>
<td>4. The Mythbusters clips helped me understand the principles of experimental methods.</td>
<td>5.8</td>
<td>1.24</td>
</tr>
<tr>
<td>5. The Mythbusters clips made class more enjoyable.</td>
<td>6.5</td>
<td>0.75</td>
</tr>
<tr>
<td>6. The Mythbusters clips helped me apply key concepts to real research studies.</td>
<td>5.9</td>
<td>1.22</td>
</tr>
<tr>
<td>7. I recommend using the Mythbusters clips again in a research methods course.</td>
<td>6.7</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Note. Responses were made on a 7-point scale with higher numbers indicating greater agreement.

and interactions. Students viewed the clip after our lecture on factorial designs.

After students viewed each clip, we distributed a handout that required them to identify the independent and dependent variables, state how the researchers operationally defined these variables, determine why the researchers chose the method that they did, and summarize the results and conclusions. The handout also included questions regarding details specific to the particular clip (e.g., “At the end of the yawn episode, Adam says ‘Given our large sample, I think it is confirmed.’ Why does the size of the sample matter?”; “In the rain episode, what extraneous variables do the researchers try to control and how do they control them?”). After receiving the handout, students met in small groups (5–7 members) and spent several minutes answering the questions. Afterward, the class as a whole discussed their answers. The length of the clips varied, and discussion typically required an additional 15 min.

Student Evaluations of the Mythbusters Clips

At the end of our course, we assessed students’ opinions about the use of the Mythbusters clips. We assured students that we would not view their responses until after we submitted the grades. Forty students answered seven questions regarding their perceptions of the clips. They made their responses to these seven items on a rating scale ranging from 1 (strongly disagree) to 7 (strongly agree). As shown in Table 2, students found the clips to be an effective teaching aid that helped them understand the course material better. Students also indicated that the Mythbusters clips helped them apply course concepts to actual research studies. Furthermore, students found the clips to be enjoyable and highly recommended their use in our future research methods courses. In fact, in the end-of-semester course evaluations, several students spontaneously mentioned that the Mythbusters clips were their favorite aspect of the course.

Exam Performance Based on Mythbusters Clips

The previous results suggest that students greatly enjoyed the Mythbusters clips and found them helpful in understanding the course material. More important, we wanted to assess if the clips were as effective as students seemed to believe. We therefore compared students’ performance on multiple-choice exam questions related to the clips with their performance on questions unrelated to the clips (for a similar procedure, see Eaton & Uskul, 2004). Students completed three exams throughout the semester: The first exam contained two questions regarding course material reflected in the toast clip; the second exam contained two questions regarding material covered in the yawning clip and two questions regarding material covered in the cell phone clip; and the third exam contained three questions regarding material covered in the rain clip. Thus, nine questions related to course material from the Mythbusters clips. We compared performance on these nine questions to performance on the 67 control exam questions (i.e., questions on material not depicted in the Mythbusters clips). Table 3 includes examples of both types of exam questions.

For each student, we calculated the percentage of items correct for the Mythbusters-relevant and control items across the three exams and then transformed these proportions using an arcsine procedure (because proportions can violate homogeneity of variance assumptions). We then used a paired samples t test to compare mean performance on the Mythbusters-relevant items with performance on the control items.
Table 3. Examples of Relevant and Control Exam Items

<table>
<thead>
<tr>
<th>Mythbusters-relevant items</th>
<th>In the Mythbusters toast-dropping episode, they created a “non-naturalistic” toast-dropping machine. The reason for this was to reduce</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(a) systematic variance</td>
</tr>
<tr>
<td></td>
<td>(b) test–retest reliability</td>
</tr>
<tr>
<td></td>
<td>(c) convergent validity</td>
</tr>
<tr>
<td></td>
<td>(d) measurement error*</td>
</tr>
<tr>
<td></td>
<td>In the Mythbusters rain episode, they created fake rain to control for a variety of extraneous variables. As a result, their study is probably</td>
</tr>
<tr>
<td></td>
<td>(a) high in internal validity*</td>
</tr>
<tr>
<td></td>
<td>(b) high in external validity</td>
</tr>
<tr>
<td></td>
<td>(c) unethical</td>
</tr>
<tr>
<td></td>
<td>(d) a correlational design</td>
</tr>
<tr>
<td>Control items</td>
<td>A study with one independent variable is called a ________ design.</td>
</tr>
<tr>
<td></td>
<td>(a) first-order</td>
</tr>
<tr>
<td></td>
<td>(b) factorial</td>
</tr>
<tr>
<td></td>
<td>(c) one-way*</td>
</tr>
<tr>
<td></td>
<td>(d) two-way</td>
</tr>
</tbody>
</table>

Scott predicts that the color blue makes us feel happy. In his study, he greets participants and places them in a blue-colored room or a white-colored room. He then has participants interact with a stranger and videotapes the interaction. Later, Scott watches the videos and records the number of times each participant smiles. Which of the following is the greatest threat to validity in this study?

(a) Pretest sensitization
(b) Experimenter expectancies* 
(c) Maturation
(d) A placebo effect

Note. Asterisks (*) indicate the correct answers.

We included data from 46 students in the analysis, eliminating data from 8 students who were present for only one of the days that the class watched the Mythbusters clips. The results indicated that the percentage of correct answers for the Mythbusters-relevant items (M = 90.3%, SD = 11.4) was significantly higher than the percentage of correct answers for the control items (M = 84.4%, SD = 10.2), t(45) = 4.9, p < .001, d = 0.7. In other words, students performed better on test items when Mythbusters clips previously illustrated the material.

Although these data suggest that watching the clips led to better performance on the Mythbusters-relevant items, an alternative explanation for our data is that the relevant items were simply less difficult than the control questions. To rule out this alternative explanation, two independent raters categorized each exam item as low, medium, or high in difficulty. Interrater reliability was high (κ = .86, p < .001). Because raters assessed all Mythbusters-relevant items as either low or medium in difficulty, we only selected the control exam items that raters had classified as low or medium in difficulty (63 items). Next, we used the same analytic procedures described earlier to compare students’ performance on these items to performance on the Mythbusters-relevant items. Once again, the results indicated that the percentage of correct answers for the Mythbusters-relevant items (M = 90.3%, SD = 11.4) was significantly higher than the percentage of correct answers for the 63 control items (M = 85.3%, SD = 10.0), t(45) = 4.6, p < .001, d = 0.5. Thus, even when all the exam items were of the same level of difficulty, students still performed better on the Mythbusters questions.

Discussion

Results suggested that instructors can use Mythbusters as an effective teaching tool in psychology research methods courses. The self-report data demonstrated that students found the Mythbusters clips to be informative and enjoyable. Moreover, the educational value of the clips was also evident in students’ actual test performance. The exam data indicated that students performed better on test questions related to concepts displayed in the Mythbusters clips than on control questions.
Results suggested that the Mythbusters clips captured students’ interest and helped them better understand key course concepts. Although we showed that the difference in performance between the Mythbusters items and the control items was not due to differences in difficulty, there are additional plausible alternative explanations for these results. First, it could be that the Mythbusters items are more concrete, because they refer to a specific example, whereas the control items are more abstract. Although this might be the case for some of our items, many of our exam questions described a concrete situation and asked students to identify a construct or concern within this example (see the last item in Table 3 for an example). In fact, 24% of the control items referred to a concrete example such as this; therefore, it is unlikely that our results are solely due to a difference in concrete versus abstract items. Second, it could be that by utilizing the Mythbusters clips, we spent more class time covering topics depicted in the clips than other topics. However, we believe it is unlikely that this explanation can fully account for our results either because we considered the time allotment issue when we first structured our lectures. Whenever we introduced a topic in class, we typically defined the term and then provided a vivid example that more clearly demonstrated the construct. However, when we introduced a topic from the Mythbusters clips, we simply defined the term, using the clip and class discussion for elaboration, devoting similar class time to Mythbusters-related topics and other topics. Thus, although differences in abstractness and time spent on topics might be plausible alternative explanations, there are valid reasons to believe that they do not fully account for our results.

Furthermore, we acknowledge that two of the clips tested myths that relied on physical rather than psychological principles (i.e., Does toast always land buttered-side down? Who gets wetter?). The use of nonpsychological examples to demonstrate research principles is a technique often seen in psychology research methods textbooks (e.g., Pelham & Blanton, 2007; Stanovich, 2004); however, some instructors might prefer to use only examples that explicitly focus on psychological constructs. We would have preferred to include only clips that test psychological myths in our study, but there were not enough choices to select from on this particular DVD. Later seasons of Mythbusters have tested additional myths related to psychology, including the use of electric shock therapy to treat seasickness (“Seasickness—Kill or Cure” episode, Season 3), the effects of Chinese water torture (“Brown Note” episode, Season 3), mind control (“Mind Control” episode, Season 4), animal learning (“Dog Myths” episode, Season 5), ways of cheating on a lie detector (“Confederate Steam Gun” episode, Season 5), and the effects of anger on driving (“Exploding Steak” episode, Season 6). We therefore encourage instructors to look for examples beyond the Mythbusters: Collection 1 DVD used in this study.

Instructors often rely on film clips in their classes because such teaching aids encourage students to go beyond simple rote memorization of concepts. Film clips encourage students to think critically about course material, and they require students to apply key concepts to real-world examples. Our study shows that psychology research methods courses can also benefit from this approach. By showing Mythbusters episodes in class, instructors can stimulate students’ excitement about research, encouraging students’ realization that, like the Mythbusters, they too can “do all their own experiments.”

References


---

**Notes**

1. The *Mythbusters* handouts used in this activity are available at the first author’s Web site: [http://psychology.okstate.edu/faculty/burkleyed/burkleyed.html](http://psychology.okstate.edu/faculty/burkleyed/burkleyed.html). To access the materials, select the “self-regulation laboratory” Web site and then select the “Publications” link.

2. Send correspondence to Edward Burkley, Oklahoma State University, Department of Psychology, 116 North Murray, Stillwater, OK 74078; e-mail: ed.burkley@okstate.edu
HOMER as an Acronym for the Scientific Method

Jessica L. Lakin\textsuperscript{a}, R. Brian Giesler\textsuperscript{b}, Kathryn A. Morris\textsuperscript{b} & Jordan R. Vosmik\textsuperscript{a}

\textsuperscript{a} Drew University
\textsuperscript{b} Butler University

Version of record first published: 05 Dec 2007
Mnemonic devices are useful tools for students who are learning new material (Balch, 2005; Carney & Levin, 1998; VanVoorhis, 2002). For example, Stalder (2005) provided several acronyms for use in introductory psychology (e.g., ABC for the three foci of psychological study: affect, behavior, and cognition) and demonstrated that students who used these memory tools performed better on exam questions about acronym-related material than students who did not use them. Students also reported that the instructor-provided acronyms encouraged them to study earlier for exams and made memorizing material easier. Stalder also found that although few instructors regularly use acronyms, they overwhelmingly believe that acronyms are helpful and acknowledge that students probably create their own. Stalder’s findings suggest that instructors who provide mnemonics in the form of acronyms will increase their students’ ability to learn and retain course information. Instructors should therefore incorporate acronyms into their coursework whenever practical.

In an effort to increase acronym use in classes beyond introductory psychology, and recognizing that research methods is typically a challenging topic for students, we developed a mnemonic strategy to help students remember the steps in the scientific method: HOMER. Each letter in HOMER stands for one of the five steps in the scientific method: hypothesize, operationalize, measure, evaluate, and replicate/revise/report (see Table 1 for sample material that we discuss at each step). The acronym engages students and provides a structure for them to remember course material. In addition, because almost all students are familiar with Homer Simpson, the television character, using this acronym connects abstract course material to a common life experience. Thus, supplementing the HOMER acronym with pictures of Homer Simpson makes the mnemonic even easier to remember. Most important, this acronym is general enough for use in any course in which instructors teach research methodology. Instructors could present the acronym in a single class period as a refresher for students who have already learned this material, or they could use it throughout a semester as a framework for students who are learning research methodology for the first time.

Evaluation of HOMER’s Effectiveness

Participants and Procedure

Forty-one students in three sections of an introductory research methods course used the HOMER

Footnote:

1Although visual images are sometimes an important part of mnemonic use, in this particular case, we used the image of Homer Simpson solely to cue the terms that comprised the mnemonic.
acronym throughout the semester. We presented HOMER as a single slide; an image of Homer Simpson appeared on the left and the words comprising the acronym, capitalized with the first letter underlined, appeared on the right. As we worked through the acronym, we highlighted the sections that we covered previously. In one of the three sections (n = 13), the instructor used HOMER intensively as a structure for the entire course (the instructor presented the acronym on 15 of 21 course days; 71.4%); in the other two sections (n = 28), a different instructor noted HOMER only when transitioning from one step of the scientific method to the next (the instructor presented the acronym on 5 of 21 course days; 23.8%).

In all three sections of the course, students saw HOMER for the last time during the second-to-last class and evaluated the acronym during the last class. First, they generated each of the five steps in the scientific method using the HOMER acronym as a guide. They then completed an open-ended question that asked them to describe in detail each of the five steps in the scientific method. Finally, they answered 11 questions about the HOMER acronym.2 Four questions evaluated enjoyment (e.g., “I enjoyed using HOMER as a framework for this class”; “To be honest, HOMER was boring”; Cronbach’s α = .77), five questions evaluated perceived usefulness (e.g., “HOMER helped me to remember the steps in the scientific method”; “HOMER made it more difficult to understand the material”; Cronbach’s α = .81), and two questions evaluated HOMER’s future usefulness (e.g., “Future research methods classes should include HOMER”; Cronbach’s α = .97). Students answered each question on a scale ranging from 1 (not at all) to 5 (extremely). We created composite measures for each question type by reverse-coding and averaging appropriate items.

Results and Discussion

On average, students correctly recalled four of the five steps in the scientific method (M = 4.17, SD = .89). Only 9 of 41 students (22%) missed two or more steps. Thus, the HOMER acronym appeared to help students remember the basic steps of the scientific method. Further, a two-sample t test revealed that students whose instructor discussed the acronym more often (M = 4.69, SD = .48) remembered the steps better than students whose instructor discussed the acronym less frequently (M = 3.93, SD = .94), t(39) = 2.75, p < .01.

The first and second authors coded the open-ended responses to determine the extent to which students accurately elaborated on each step in the scientific method. Interjudge reliability was high, r(37) = .83, p < .01, so we averaged the two judges’ ratings. Scores could range from 0 (elaboration on none of the five steps) to 5 (elaboration on all five steps). Students elaborated successfully on most steps (M = 3.33, SD = 1.13). A one-sample t test demonstrated that students’ understanding of the scientific method was significantly higher than 0, t(37) = 18.18, p < .01. Again, a two-sample t test revealed that students whose instructor discussed the acronym more often (M = 3.96, SD = 1.12) elaborated on the steps more successfully than students whose instructor discussed the acronym less frequently (M = 3.04, SD = 1.03), t(36) = 2.50, p = .02. Thus, use of the HOMER acronym helped students remember specific material about the scientific method.

Overall, students also reported that using HOMER was enjoyable (M = 3.70, SD = .76) and useful.

Table 1. Sample Material Instructors May Teach at Each Stage of HOMER

<table>
<thead>
<tr>
<th>Step</th>
<th>Material</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothesize</td>
<td>Characteristics of the scientific method, development of research ideas,</td>
</tr>
<tr>
<td>Operationalize</td>
<td>Translation of conceptual variables into operational definitions, scales</td>
</tr>
<tr>
<td>Measure</td>
<td>The three basic types of research (descriptive, correlational, experimental) and the characteristics of each, variations on the three basic research designs (e.g., quasi-experiments, factorial designs)</td>
</tr>
<tr>
<td>Evaluate</td>
<td>Appropriate statistical analyses for each type of research design, internal and external validity, statistical validity and Type I and II errors, power</td>
</tr>
<tr>
<td>Replicate/revise/report</td>
<td>Types of replications, importance of revisions, oral presentations, conference-style poster presentations, APA-style written reports</td>
</tr>
</tbody>
</table>

2Degrees of freedom for the reported analyses differ slightly due to incomplete assessment data.
Table 2. Two-Sample t Tests Comparing Effect of Frequency of Use of HOMER on Enjoyment, Perceived Usefulness, and Future Use

<table>
<thead>
<tr>
<th>Frequency of Use of HOMER</th>
<th>Intensive (15 of 21 Days)</th>
<th>Nonintensive (5 of 21 Days)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Enjoyment</td>
<td>4.25</td>
<td>.42</td>
</tr>
<tr>
<td>Perceived usefulness</td>
<td>3.82</td>
<td>.76</td>
</tr>
<tr>
<td>Future use</td>
<td>4.12</td>
<td>.82</td>
</tr>
</tbody>
</table>

Note. Scores could range from 1 (not at all) to 5 (extremely). All tests conducted with df = 38.

(M = 3.29, SD = .74) and noted that instructors should use it in future courses (M = 3.54, SD = .90). For all three question types, one-sample t tests revealed that students rated HOMER significantly higher than 3, the midpoint of the evaluation scale, ts(39) = 5.81, 2.42, and 3.77, respectively, all ps < .02. The benefits of HOMER were more apparent to students whose instructor discussed it more often. Table 2 reveals that students whose instructor discussed HOMER more often were more likely to report that HOMER was enjoyable, useful, and important to include in future classes.

Stalder (2005) provided evidence of the usefulness of acronyms as mnemonic strategies for students in introductory psychology. We add to this literature by demonstrating the effectiveness of a fun acronym for learning about the scientific method. Even though the evaluation of the acronym occurred in classes with two different instructors and we did not randomly assign students to sections of the course, the overall learning that occurred and the positive evaluation of the acronym suggest it will be a beneficial tool to instructors teaching research methodology. Although we evaluated the efficacy of this acronym in research methods courses, we have used it in other courses, including introductory psychology, social psychology, and psychology of gender. Further, we have used the acronym in courses ranging in size from 10 to 100 students. In summary, we believe the acronym would be an effective addition to any course in which instructors teach the scientific method, and we hope that other instructors will find this acronym to be a useful addition to their courses.

References


Notes

1. Support for this research was provided by a release time grant from Drew University to Jessica L. Lakin.
2. We are grateful to Randolph A. Smith and three anonymous reviewers for their helpful comments on an earlier version of this article.
3. Send correspondence or requests for a complete list of dependent measures to Jessica L. Lakin, Psychology Department, Drew University, 36 Madison Avenue, Madison, NJ 07940; e-mail: jlakin@drew.edu.
Using Single-Case Design and Personalized Behavior Change Projects to Teach Research Methods

David L. Morgan
Spalding University

Students in research methods courses, especially those taught in an intense format, might be hard pressed to actually conduct research studies due to logistics and time constraints. I describe the use of single-case research design and a personalized behavior project as an alternative research project for students in an undergraduate psychology methods course. Students were able to complete the project, including the collection, analysis, and interpretation of data, within the confines of an intensive, 6-week course.

Research methods courses can be among the most challenging to teach to psychology majors, in part because these students’ interests in psychology tend to run toward clinical and applied areas, not empirical research (Rottinghaus, Gaffey, Borgen, & Ralston, 2006). In addition to the technical and often overwhelming nature of the content, students are often unable to actually experience the research process within the temporal restraints of an academic semester, particularly in intensive courses. Several years ago, Spalding University (Louisville, KY) reformed its undergraduate programs, initiating an academic format in which students take classes in intense 6-week (100 min × 4 days a week) sessions. This new, intense format rendered the objective of having students conduct actual research projects untenable, eliminating the opportunity for a hands-on research experience.

The logistics of actually conducting research within a classroom context have always been somewhat daunting, with major constraints being the requirement to secure institutional review board approval prior to data collection and the substantial sizes of participant samples needed to allow for meaningful statistical analyses. At Spalding, the 6-week session format, coupled with the historical obstacles of conducting research, left little room for students in research methods courses to experience research firsthand.

Fortunately, recent years have seen an upsurge in methodological innovation, and alternative strategies for conducting empirical research, even within the constraints of an intensive academic course, are available to faculty teaching methods courses. One such innovation is the single-case research design, also known variously as single-case, small-n, or N = 1 research design. Single-case research design actually has a long and storied history within psychology (e.g., Kazdin, 1982; Morgan & Morgan, 2001; Sidman, 1960), but is perhaps less familiar to many university faculty. Single-case research boasts several advantages, however, over conventional null-hypothesis testing group designs. Among these advantages are the small participant samples required to conduct such research and repeated measurement of dependent variables in real time. These design features encourage a fine-tuned analysis of behavior both prior to and following independent variable manipulations. Large-group, statistically driven studies relying on aggregated summary measures seldom accommodate the degree of measurement resolution typical of single-case studies.

This project involved introducing students in a sophomore-level research methods course to single-case design, with the requirement that students conduct a personalized research project aimed at bringing about change in a personally relevant behavior. Students chose a simple, operationally defined target behavior that was easy to observe and record, such as exercise, caffeine consumption, or study time. For the first 2 weeks of the academic session, students collected baseline measures of the target behavior and
plotted these data on a standard time-series (real-time) graph. Students then devised their personal intervention strategies for changing the target behavior in a desirable direction. For the next 10 to 14 days, students collected and plotted data on the target behavior during this intervention phase. Finally, students conducted an analysis and interpretation of the data from their personalized project, using statistical process control (SPC) procedures (Hart & Hart, 2002).

SPC analysis, and control charts in general, are common strategies within manufacturing to evaluate and maintain quality control of products. Shewhart (1931) introduced the method to manufacturing in the 1920s, and SPC analysis eventually became the gold standard for evaluating product quality, initiating the total quality management revolution (Berman, 1995). In recent years, behavioral scientists and health care professionals have argued the merits of SPC for evaluating behavioral data, including clinical interventions (Callahan & Barisa, 2005; Carey & Lloyd, 1995; Orme & Cox, 2001; Pfadt & Wheeler, 1995).

As is often the case, many students chose target behaviors they considered to have health consequences, including exercise and dietary habits. Actual data from one student's project, plotted on a time-series graph with SPC analysis, appear in Figure 1. The graph depicts the number of calories consumed by snacking in the evening following dinner. For this student, evening snacking was undesirable and the goal of the project was to reduce this postmeal consumption. Horizontal lines running through both the baseline and intervention phases of the graph depict the mean evening calories consumed during baseline (solid line) and a 2 SD band or confidence interval (referred to as upper and lower control limits in SPC language) about this mean. In SPC analysis, control limits calculated on baseline data serve as criteria for interpreting behavior change during the intervention phase. Data that fall outside the control limits, as evident especially in later portions of the intervention phase in Figure 1, indicate statistically rare, or “special cause,” data. Such deviations from baseline responding are usually interpreted as evidence of an independent variable or intervention effect. The results depicted in Figure 1 are fairly typical of those collected by students in the current class project.

Students completed a 10-item multiple-choice quiz (pretest) specific to single-case research design and statistical process control during the first day of class. Most of the instruction regarding single-case epistemology and data collection, including discussion of how to plot data on a time-series graph, occurred late in the session, as single-case research is among the last chapters in the text used in this class (Leary, 2008). On the last day of class, students completed the same quiz (posttest) for a second time. Although data for only 10 students were available for both pretest ($M = 6.8$) and posttest ($M = 7.8$), the results of a dependent $t$ test indicated a significant increase in mastery of single-case research design and statistical process control, $t(9) = -3.35$, $p < .004$, $d = .70$.

The single-case behavior change project recommends itself to instructors of research methods courses for several reasons. First, students are necessarily actively involved in their project, a variable known to enhance learning (Menges & Dobroski, 1977; Yoder & Hochevar, 2005). Second, because students collect
personally relevant data, their attention to the project and their interest in the actual data generated are nearly guaranteed. Third, the project is quite manageable, even within the considerable constraints of the intensive format, because students are able to accomplish all phases of data collection and analysis within the necessary time frame. Because students are collecting personal data not intended for public dissemination, no institutional review board review is necessary. Consequently, students can identify a target behavior and begin data collection soon after they receive the assignment. Fourth, SPC analysis allows students to use predetermined quantitative criteria, rather than personal judgment, to evaluate behavior change. Perhaps most important, the project allows for an intimate, hands-on experience, making the scientific enterprise less amorphous and mysterious to the student.

Although students in the course studied here conducted their behavior change projects in a short, intensive format, longer academic sessions would afford students an opportunity to explore additional features of single-case designs. Single-case designs are experimental designs, and like other experimental strategies, they vary in their ability to control for threats to internal validity. With additional course time, students could collect data across several phase changes, allowing for potential replication of the treatment effect. Additional project phases also set the stage for discussions of the value of replication in research, and the various advantages and disadvantages of single-case design alternatives, such as reversal, multiple-baseline, and changing criterion designs (Morgan & Morgan, 2009). Finally, students can be encouraged to compare the results of their SPC analyses to alternative analyses. Numerous strategies for analyzing single-case data exist, some having been introduced quite recently. Some of these analyses require fairly simple calculations and lead to intuitive interpretation, and students engaged in a comparative exercise would be making contact with state-of-the-art developments in the field of single-case design and analysis (Fisher, Kelley, & Lomas, 2003; Parker & Brossart, 2003; Parker & Hagan-Burke, 2007).

References


Note

Send correspondence to David L. Morgan, Spalding University, 845 South Third Street, Louisville, KY 40203; e-mail: dmorgan@spalding.edu
Using Community-Based Research Projects to Teach Research Methods
Andrea Chapdelaine and Barbara L. Chapman
Teaching of Psychology 1999 26: 101
DOI: 10.1207/s15328023top2602_4

The online version of this article can be found at:
http://top.sagepub.com/content/26/2/101

Published by:
http://www.sagepublications.com

On behalf of:
Society for the Teaching of Psychology

Additional services and information for Teaching of Psychology can be found at:

Email Alerts: http://top.sagepub.com/cgi/alerts
Subscriptions: http://top.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Apr 1, 1999
What is This?

Notes
1. We thank the anonymous reviewers for comments on an earlier version of this article.
2. Send correspondence to James E. Hartley, Department of Economics, Mount Holyoke College, South Hadley, MA 01075; e-mail: jhartley@mtholyoke.edu.

Using Community-Based Research Projects to Teach Research Methods

Andrea Chapdelaine
Barbara L. Chapman
Trinity College

In this article, we describe the elements comprising a community-based research project conducted as part of an undergraduate research methods course and the project’s positive impact on student learning. A community-based project provides a unique opportunity for students to learn the methodology of psychology, directly experience the research process, and make a significant societal contribution.

Most psychologists recognize the importance of involvement in the research process as a means to learn the scientific method, a view that researchers have supported empirically (Conrad & Hedin, 1991; Lehman & Nisbett, 1990; VanderStoep & Shaughnessy, 1997). Accordingly, undergraduate students may assist faculty with their research or conduct independent research projects (Durso, 1997; Kierny, 1984). However, these research experiences often occur late in the undergraduate career. We believe that students should become involved in research in the introductory research methods course. Such early exposure will prepare students to take advanced courses in the discipline, to serve as research assistants, and to pursue an independent senior project. Including hands-on research experience in a methods course has demonstrated benefits, including improved understanding of methodological concepts, enhanced critical thinking skills, and greater interest in course material (Durso, 1997; Evans, Rinta, Guthrie, & Raines, 1981; Markus, Howard, & King, 1993).

However, given time constraints, large enrollments, and students’ lack of research expertise, including research is a difficult task for the instructor. Instructors have developed course research projects in ways that address these problems, most often by circumventing one of the steps in the research process. For example, students may analyze existing data sets, propose rather than conduct research, perform computer-simulated experiments, or conduct group projects. Although these solutions alleviate some of the obstacles to
conducting research in the methods course, each comes with its own problems. Students may miss the opportunity to design a study or collect data, be unable to complete the project within one semester, or produce results unsupportive of the hypothesis, leading to disappointment and disengagement.

Recently, there has been renewed interest in including community service as part of the college curriculum (Markus et al., 1993; Mettetal & Bryant, 1996; Rubin, 1990). Conrad and Hedin (1991) reported that community service increases factual knowledge of a topic area, self-esteem, concern for others, and awareness of one's civic responsibilities. Community service internships, commensurate with the values of psychology as a discipline, have been a long-standing component of the undergraduate psychology curriculum (Raupe & Cohen, 1992). However, internships often are considered distinct from research engagement, which traditionally occurs in on-campus faculty laboratories (Herzberger & Chapdelaine, 1998).

A community-based research project is applied research (e.g., Cook & Campbell, 1979; Posavac, 1997) that provides needed information for a community and a learning opportunity for the students working on the project (Mettetal & Bryant, 1996). We believe that this type of research project overcomes some difficulties of providing research experience within the constraints of a methods course. First, although the course instructor determines the topic, the project is broad enough to allow students to pursue their own interests within the project and, due to the project’s social value, is likely to be of interest to students. Second, except for instrument design, students execute all aspects of the research process. Third, the instructor’s supervisory duties are focused on a single project, and thus, he or she can exercise more control over the research. Moreover, the project integrates community service with research aspects of the psychology curriculum.

As constructors of a methods course, we designed a community-based project in conjunction with the local police department. The department had initiated a domestic violence program, the goals of which were to increase early intervention by police, services for victims, successful prosecution of perpetrators, and public education on domestic violence (Hartford Police Department, 1995). As part of this program, the class surveyed residents’ attitudes toward the role of the police in handling domestic violence cases. The police department used the results to develop community education and officer training programs as well as to provide baseline statistical information to the funding agency. Following is a discussion of each phase of the project in terms of its execution and relation to the course material and student learning.

Project Design and Implementation

Selection and Design of Project

Selecting the project is one of the hardest and yet most critical steps. We were fortunate that there is an initiative at our institution to utilize the city as an educational resource, and thus, we received assistance in identifying an appropriate project. Regardless of how contact is made, there are important issues to bear in mind when selecting a project. Accord-

ing to Mettetal and Bryant (1996), the project must relate substantially to the course goals and must provide a necessary service for the community. On a more pragmatic level, we would add that the selected agency must be well-organized and fully informed of the parameters of the project. The agency must be willing and able to provide the students with necessary support, supervision, and information; must have a clear plan for utilizing the results; and must be able to adhere to the time schedule. Furthermore, the agency must fully endorse the students’ efforts and the study’s results must be worth the agency’s investment. To address these issues, ample communication between the instructor and the community representatives must occur well in advance of the start date and throughout the course of the project.

Project Planning

Prior to the semester, in conjunction with police officers overseeing the project and course teaching assistants (undergraduate students who previously took the course), we designed a survey, selected a data collection method, and discussed police use of the results. To include a larger number of residents, we selected a phone survey as the data collection method. Moreover, due to the domestic violence topic, the class sampled only women because the questions might have aroused suspicion in male respondents who were batterers. We believed students were not trained to handle such a potentially difficult situation, and we did not want to inadvertently cause a battering event by raising such suspicions. Finally, we generated a task list that served as a template for course assignment deadlines (e.g., hypothesis selection, final report).

This first phase of the project was the only one in which the students were not directly involved. During the semester however, students critiqued the survey for potential sources of bias and validity threats and generated additional or revised questions. Similarly, while reviewing data collection methods, students discussed the advantages and disadvantages of a telephone survey and the possible consequences of using this method on the project’s results. These class discussions allowed students the opportunity to be involved, albeit in a post hoc manner, by applying corresponding course material to the initial phase of the project.

Developing a Hypothesis

As part of course requirements, students wrote an independent report on a hypothesis they formulated and tested. We took several steps to provide students with the background information necessary to generate a testable hypothesis. First, the class discussed the police program (Hartford Police Department, 1995), the survey instrument, and a book on domestic violence (Herzberger, 1996). Second, two police officers visited the class to provide information on the project. They also suggested questions the students could explore that would be useful with regard to the program objectives. This contact with the community agency was important, as it made concrete for the students that the project was “real” and catalyzed their interest. Third, the students wrote a critique of the back-
ground reading and presented potential hypotheses. Fourth, the authors wrote a brief guide on how to generate hypotheses that included possible independent and dependent variables. Finally, each student met with one of the authors to discuss his or her ideas and solidify a hypothesis. A helpful resource for students was a set of reserve readings and a corresponding annotated bibliography. Although students used these references, they also had to include references they found independently. Thus, students learned how to use the library databases and conduct a literature review.

Data Collection

This project’s findings were important to the police and the funding agency; thus, it was critical to control data collection strictly. As the first step in data collection, students generated the random phone numbers that they subsequently called. Specifically, we gave students sections of the local phone book, a random start number, and log sheets for recording the numbers. This exercise provided an excellent lesson on random selection.

Training for survey administration involved several steps. First, students spent a laboratory period discussing the survey administration procedure, including how to ask the questions, code responses, and answer questions about the project. Also, we gave students a handout on “Do’s and Don’ts for Conducting Interviews” (Backstrom & Hursh-Cesar, 1981). Second, each student called a teaching assistant prior to data collection. The teaching assistant role played a respondent and purposely raised issues that might pose difficulty for the student interviewer (e.g., how to politely steer a respondent back to answering a question). At the end of the call, the assistant discussed areas of strength or weakness and, where needed or requested, scheduled a second call. Third, an instructor or teaching assistant was present at each data collection period to monitor students. Students were concerned about being a source of bias in the project and thus were well prepared for data collection. Students collected data during weekdays or evenings over a 1-week period.

Data Entry

As with the data collection, it was important to ensure that the data entry was correct. We improved accuracy by providing a detailed codebook, supervising data entry, having students work in pairs so they could monitor each other’s work, having the teaching assistants check a random subset of the data, and grading based on precision. Data entry introduced students to the statistical computer program they used to analyze their data. Because data entry required an understanding of only the basic commands, this assignment was an optimal way to begin learning the program.

Hypothesis Testing

Students analyzed their results in two stages. First, students generated descriptive data for the variables they used in their analyses. These initial analyses enabled students to obtain a sense of what the data “looked like.” Second, we formed small groups of students according to the similarity of their hypotheses. These groups met with one of the authors to discuss programming and interpretation of the results. In addition, groups collaborated on the data analysis, which was an opportunity for students to model the way we, as their mentors, conduct research. Although this collaboration was possible due to some overlap in the students’ hypotheses, each hypothesis was unique enough that every student did some independent statistical work.

Project Reports

A primary goal of many methods courses is to teach students how to write an empirical report. Students’ reports constituted a majority of their project grade. To provide ample feedback on their writing, we divided this project component into several deadlines throughout the semester, such that students produced a draft and revision for each report section. In addition, students reviewed each others’ drafts, which demonstrated the importance of constructive feedback in the writing process. This peer review was feasible because all students worked on the same project and thus critiqued a paper on a topic with which they were familiar. Students also became more aware of the dangers of “writer-based prose” (Tate & Corbett, 1981, p. 270), in which authors leave much unexplained or assumed. At the end of the semester, students incorporated their revisions into a final report. Some students presented their findings at the college science fair, some at a local undergraduate research conference, and others prepared their manuscripts for publication.

Presentation to Community Agency

Students submitted a brief summary of their findings and the findings’ implications with regard to the program goals to the police. For example, one student found that although residents supported police intervention in domestic violence, the type of intervention most supported for minor abuse (i.e., slapping) was counseling rather than prosecution. The state has a mandatory arrest policy, so the student discussed how the police might shape the public education component of the program around this issue. In addition, the police could request the students’ complete reports. Although a poster or oral presentation to the police was not feasible in this project, we strongly encourage either presentation format to provide students the opportunity to discuss their findings with members of the community agency.

Project Evaluation

The primary purpose of this project was to provide students enrolled in an introductory methods course the opportunity to directly experience the research process early in the undergraduate psychology curriculum. Such active learning fosters mastery of course material (McNamara & Healy, 1995). On the course evaluations, students indicated the ex-
tent to which the project enhanced their learning of the course content on a 5-point scale ranging from 1 (it detracted greatly from my learning and understanding) to 5 (it added greatly to my learning and understanding). Of the 25 students, all but 2 gave a rating of 4 or 5 to this question, resulting in a mean rating of 4.32 (SD = 0.75).

Further evidence of the project’s utility to course learning were the ample opportunities it provided students to apply methodological concepts and issues to the project. In addition to those noted previously, one engaging application was the discussion of research ethics. During training for data collection, students discussed what they should do if a respondent indicated that she was abused. Students provided numbers to a domestic abuse hotline and nothing more, as they were not trained counselors. We emphasized that this procedure was consistent with ethical guidelines, was approved by the institution’s review board, and was in accordance with the need to maintain objectivity in one’s research. Although students understood this reasoning and followed the prescribed guidelines, this discussion was an important lesson on the tension between objectivity and sympathy. Similarly, the class discussed the decision to exclude male respondents, as such exclusion limited the findings’ generalizability. Students weighed the potential costs to participants (and their partners) and the benefits of the research and learned how ethical decisions are not always easy or apparent.

In addition to ethical issues, the project provided examples of many other course concepts, from experimenter expectancy effects to a demonstration of creating graphs in ways that seriously distort the results of a study. In fact, students often raised issues and applied material on their own. As evidenced by their grades on successive writing components, students’ writing improved considerably over the course of the semester. In sum, there was strong evidence that the project enhanced critical thinking, understanding of methodological concepts, scientific writing, and reasoning skills.

A secondary goal for using this project was to enhance student interest in and enjoyment of research (Evans et al., 1981). Students rated how positive they found the project independent of its connection to the course. Ratings made on a 5-point scale ranging from 1 (very negative) to 5 (very positive), indicated that students found the experience highly positive (M = 4.4, SD = 0.5). As one student stated, “the project made the course material ‘real-life’ and ‘interesting.’” Another wrote: “It was great to have hands-on experience.”

A final goal of the project was to help students understand the importance of research in a larger context by participating in a project that would impact the lives of people residing in the community surrounding the campus. As others have argued, education should involve teaching civic responsibility (Raupp & Cohen, 1992; Rubin, 1990). Students attested to increased awareness of social issues. Research shows that social awareness encourages a sense of responsibility and social empathy (Conrad & Hedin, 1991).

**Limitations and Future Considerations**

Although students were involved in the major steps of the research process, we had to complete some aspects before the semester began, including choice of topic and instrument. Students had no voice in the selection of the project, but we compensated for this problem by having students develop their own hypotheses and operational definitions for their individual reports. The topic of police intervention in domestic abuse is broad and the instrument comprehensive enough that every student could test a different hypothesis, ranging from ethnic differences in approval of abuse to the disparity between citizens’ opinions and police department policies. Thus, it is important to choose a topic that allows a wide array of hypotheses and integration of a variety of interests (e.g., psychology of the law, industrial/organizational psychology, clinical psychology).

Although students did not design the survey, we attempted to give the students a sense of ownership by devoting substantial class time to discussing the survey items. Although students knew they could not change the survey for their immediate project, they discussed rewording or adding items for a planned posttest. These discussions demonstrated the difficulty of developing an instrument, as students’ proposed changes often created new problems of wording and experimenter bias.

One problem with a community-based research project is the use of only one methodology (e.g., phone interviews) in the course. Either developing several smaller scale projects or conducting multimethod studies on a single topic might create a more thorough learning experience. For example, students also could have conducted archival research on the incidence of domestic abuse.

In this type of project, the instructor makes a huge investment of time, not only to the administration of the project in the class, but also to the agency. As principal investigators we needed to ensure the accuracy of the results prior to presenting them to the police. We served as consultants, helping the agency understand findings and answering additional questions that the students’ reports did not address. In other words, this project was a community service opportunity for the instructors as well as for the students, with substantial intrinsic benefits, including acknowledgment of the service by the college and community, the beneficial nature of the research, and rewarding interactions with students and the community.

The time commitment required of the students is also significant. As indicated by course evaluations, the students would rather have the extra work than not have the research experience. Ideally, this project should occur in a laboratory methods course, with students receiving additional credit for the laboratory. A statistics course should be a prerequisite for the course. Teaching assistants or a laboratory instructor are essential.

Unlike canned classroom demonstrations, a community-based research project often involves aspects that are beyond the instructor’s control, including the number of respondents in the community who agree to participate and the introduction of error by multiple student experimenters. Also, unlike the results of a contrived laboratory exercise, which are often discarded, the community organization uses these results; thus, they must be sound. Contingency plans may be necessary; for example, bilingual research assistants interviewed respondents from Spanish-speaking households.
The impact of the project on the community may not be seen until long after the course ends. We maintained an electronic mail list of the students so that we could share news regarding the program’s progress. Finally, the community agencies that are in need of “free” research often are those representing the disadvantaged members of our society. Exposure to these persons may be a novel experience for many college students. With this experience comes fear, insecurity, shame, and guilt (Conrad & Hedin, 1991; Raupp & Cohen, 1992). Instructors must be prepared for these emotions and help students see them as part of the learning experience. Contact with disadvantaged individuals can be an empowering experience (Mettetal & Bryant, 1996), as evidenced by the fact that several students who participated in the police intervention study now volunteer at a local domestic abuse shelter.

Conclusions

Although using community-based research projects in research methods courses can be time and energy consuming, the rewards are great. Students appreciate the experience and gain much more than a basic understanding of methodological concepts. The project is a unique opportunity for them to collaborate with community agencies and to achieve tangible goals.

References


Notes

1. Portions of this article were presented at the Fourth Northeastern Conference for Teachers of Psychology, Ithaca, NY, October 1997.
2. We thank Sharon Herberger, the Hartford Police Department, Trinity Center for Neighborhoods, teaching assistants Kerry Bachman and Monica Pernal, the students for their invaluable contributions to the project, and anonymous reviewers for comments on an earlier version of this article.
3. Class handouts mentioned in the article are available to interested readers on request.
4. Send correspondence to Andrea Chapdelaine, Department of Psychology, Albright College, P.O. Box 15234, Reading, PA 19612–5234; e-mail: andreadc@joe.alb.edu.
Teaching Observational Research in Introductory Psychology: Computerized and Lecture-Based Methods

Victoria A. Kazmerski
Dawn G. Blasko
School of Humanities and Social Sciences
Penn State Erie, The Behrend College

In the introductory psychology course, one major goal is to teach that psychology is a science. Course projects that involve data collection can be a way to engage students and to foster an appreciation of the scientific method. Courseware for Observational Research (COR) is a multimedia program designed to teach basic methods of observational research. Students in 2 introductory psychology classes conducted observational research projects; 1 class was taught using traditional methods and the other using the “Lessons” component of COR. The class taught using COR showed higher performance on a laboratory-based research report. The 2 classes performed similarly on multiple-choice questions concerning the basic concepts. Both classes reported enjoying the project; however, COR appeared to enhance students’ perceptions of the importance of research.

Instructors report that two of their most important goals in the introductory psychology course are to engage students in scientific inquiry of psychological processes and to convince students that psychology is a science (Miller & Gentile, 1998). Having students conduct their own observations can be an effective way to use active and collaborative methods to meet these goals. In this article, we describe a study conducted in our introductory psychology classes using observational research as a technique to foster learning about the scientific method. We compared teaching observational research using a standard lecture format and using an interactive software program that we originally developed for the basic research methods class.

Computer-based teaching methods have become fairly common in psychology. Forsyth and Archer (1997) reported that such methods may improve knowledge transfer as well as increase enthusiasm and motivation for learning. Using these methods for teaching research methods has become common. Several commercial programs are available to demonstrate or teach experimental research methods to undergraduate psychology students. However, there is a dearth of programs available for teaching observational research. Courseware for Observational Research (COR) is a multimedia program designed to fill this void (Blasko, Kazmerski, Corty, & Kallgren, 1998).

COR is organized into three sections: “lessons” is used by an instructor during class, a “laboratory” section is used independently by students to apply to concepts learned in class, and a “library.” The lessons section uses digitized video to illustrate three basic coding strategies: frequency coding, duration coding, and interval coding. It also provides instruction on how to simplify complex coding situations through the use of event, time, and individual sampling. Lessons are also included for assessing interrater reliability using percentage agreement and Cohen’s kappa and for assessing statistical significance using the chi-square goodness-of-fit test (for more detail on lessons, see Blasko et al., 1998). The laboratory section provides five video clips that students can use to conduct observational research. The program guides students through each of the steps in the lessons and allows them to link back to the lessons to review the concepts. The library provides direct access to the videos, background information on the video subjects, and interviews with the caretakers of some of the video subjects.

We designed the program to improve on traditional lecture-based instruction in a number of ways. First, COR uses video clips carefully chosen to illustrate the steps of observational research methods. It goes beyond simply showing videos by presenting the video clips along with interactive coding sheets that instructors can use in at least three ways. The videos can be shown alone, allowing the students to observe behaviors and develop operational definitions. They can also be shown with sample coding sheets that students can copy and code at their desk or one student can code on the computer in front of the class. In addition, the video can be played while the computer fills in the coding sheet based on data the instructor enters before class. The second benefit of COR is that it helps students by taking them through the steps of calculating interrater reliability and chi-square. Instructors or students can input data directly into the formulas and see the output at each step of the calculation. This feature is particularly useful for the math-phobic student. Therefore, COR’s primary benefit is the flexibility of use by the instructor and the precision of the presentation of concepts to the students.

To test the efficacy of the lessons component of COR, we conducted a study with two sections of introductory psychology students. One section was taught observational research using COR and the other with traditional lecture-based methods. A guest lecturer taught the unit to the two sections so that the regular class instructors, who were the authors...
and developers of COR, would not bias the results. Evaluation was done on three levels: a multiple-choice exam tested the key concepts in the unit, a research project where pairs of students conducted a brief hypothesis-driven naturalistic observation and wrote a research report on it, and an evaluation survey asked students’ opinions of the experience.

Method

Participants

Students in two introductory psychology classes (n = 44; n = 41) at Penn State Erie participated as part of their course grade. Forty-nine percent were women. Students in the two classes had similar SAT scores. The mean verbal SAT score for the first class was 531 and for the second class was 537, t(69) = 0.34, p > .05. The mean math SAT score for the first class was 546 and for the second class was 538, t(69) = 0.67, p > .05.

Procedure

The two classes met in the same classroom 2 days per week for consecutive 75-min class periods. The sections were normally taught separately by two instructors (the authors), who for the purpose of the study collaborated to use the same syllabus, texts, and course projects for the two classes. A guest lecturer, who did not participate in the development of COR, instructed the two classes on the topic of observational research. The instructors introduced the guest lecturer as an expert in the field of observational research because she often uses these methods in her own developmental psychology research. A flip of a coin determined the method of instruction for each class; the guest lecturer used the traditional lecture method (lecture-instructed class) for the first class and used COR for second class (COR-instructed class).

The students in both sections were informed at the beginning of the semester that they would be taking part in a special project. They were told they would conduct their own research project rather than take part as research participants as is traditional in the introductory psychology class. The research project was scheduled to begin 3 weeks after the semester began and the students had 4 weeks to conduct their observation. In addition to the chapter on research methods in their textbook, the students read a section on naturalistic observation from a research methods textbook (Bordens & Abbott, 1996).

The regular instructors first taught the basics of research and then the scientific method over two 75-min classes. The guest lecturer taught the next two regularly scheduled class periods. In both classes she covered the concepts of hypothesis testing, operational definitions, coding schemes (frequency, duration, and interval coding), sampling techniques with a focus on individual and event sampling, and interrater reliability using percentage agreement and Cohen’s kappa. She concluded the unit with a discussion of significance testing using the chi-square goodness-of-fit test.

In the lecture-instructed class the guest lecturer explained the concepts verbally and used overhead transparencies. She also used video clips of child participants from her research to help illustrate the concept of operational definitions. In the COR-instructed class, she used the lessons component of COR on a projection monitor. In both classes students practiced these concepts in small groups in class using pencil and paper. Course instructors were not present during the classes.

After the lessons, both classes received a handout describing the research assignment. They completed the assignment outside of class time. Students worked in pairs to develop a hypothesis for an observation or chose one of the sample hypotheses given by the instructors (e.g., “People who drive sports cars are less likely to come to a complete stop at a stop sign then those who drive other cars.”) They operationally defined their variables and conducted the observation for at least 30 min. They assessed interrater reliability of their observations using both percentage agreement and Cohen’s kappa. They tested their hypothesis using chi-square. Finally, each research team wrote up their results in a simplified American Psychological Association-format research report. Students did not need to use the laboratory component of COR to complete the assignment.

Evaluation Methods

Multiple-choice exam. Both classes answered the same 10 multiple-choice questions that asked about basic concepts and definitions in observational research as part of their first exam. All of the tested concepts were in the assigned readings.

Student research reports. To make an objective assessment, we removed the students’ names from each paper and mixed the papers from the two classes. We graded the research reports using a checklist of specific criteria. The criteria included accurate title, statement of the problem investigated, clearly stated hypothesis, operational definition of the variables, explanation of the coding scheme used, description of the sampling technique used, presentation of the interrater reliability and chi-square results, summary of the data in a table, valid conclusions based on the data, and accurate format. We graded three reports together to establish a standard grading criteria. We then graded five papers using the criteria to establish reliability. We agreed on the grades of each of the five test papers to within 2 points out of 100. We then randomly divided the remaining papers for grading.

Student evaluations. The classes rated the unit using two questions. The first question was “How useful do you think it is to learn about research by actually conducting an observational study as opposed to simply reading about it?”, rated on a scale of 1 (not at all useful) to 10 (extremely useful). The second question was “How well did the guest lecturer teach you about the purpose and process of observational research?”, rated on a scale of 1 (not at all) to 10 (extremely well). Students were also able to provide open-ended comments.

Results

Multiple-Choice Exam

On the 10 multiple-choice questions the lecture-instructed class obtained a mean score of 76% (SD =

Teaching of Psychology
12.7) and the COR-instructed class obtained a mean score of 78% (SD = 15.03). These scores did not differ reliably, \(t(82) = 0.51, p > 0.05\).

**Student Research Reports**

The COR-instructed class earned higher grades (\(M = 86\%\), SD = 11.75) than the lecture-instructed class (\(M = 79\%\), SD = 10.91) on the research reports. This difference was statistically reliable, \(t(81) = 2.80, p < 0.01, \omega^2 = 0.08\).

**Student Ratings**

Both groups of students were positive about conducting their own research in the class (\(M = 7.62\) on a scale of 1 [not at all useful] to 10 [extremely useful], SD = 2.03). In the final course evaluations, several mentioned the project as the best part of the class. One senior commented that this was the first general education course that had provided a challenge. The few negative comments about the project focused primarily on having to work in groups, with difficulty coordinating schedules or unequal distribution of work between partners. In comparing the two teaching methods, the COR-instructed class rated the usefulness of conducting research more highly (\(M = 8.2\), SD = 1.96) than the lecture-instructed class (\(M = 7.1\), SD = 2.29), \(t(71) = 2.17, p < .05, \omega^2 = .05\). The two classes did not differ, \(t(69) = 0.83, p > .05\), in how they rated the guest lecturer; on a scale of 1 (not at all) to 10 (extremely well) the mean rating for the COR-instructed class was 5.8 (SD = 2.16) and the mean rating for the lecture-instructed class was 6.2 (SD = 2.39). On reflection, the guest lecturer reported spending somewhat more time working with students in small groups practicing the concepts in the lecture-instructed class.

**Discussion**

We draw two conclusions from these results. First, asking students to complete their own psychological research is an excellent way to illustrate how the scientific method is central to psychology. Although certainly more time consuming to develop and evaluate than multiple-choice exams, these projects actively engage students to think about alternative hypotheses and consider multiple explanations of their results, some of the hallmarks of critical thinking.

Second, although multimedia programs like COR are not necessary to meet these goals, they provide a way to capture students’ attention and clarify critical concepts. Multimedia presentations offer the advantages of presenting information in an incremental manner. They offer the instructor greater precision and control and allow him or her to customize learning to the individual group of students (Goolkasian, 1996). COR offers the additional benefit of being highly interactive. The instructor can demonstrate a variety of possible outcomes using the class’s actual data without additional preparation time. For example, one of the most difficult concepts to teach is the importance of precise operational definitions. Using COR we have found it useful to show a video (e.g., a gorilla) and have the students pick a behavior (e.g., feeding) to code. The class watches the video and records the number of behaviors. When we compare each student’s response, there is wide disagreement. We show the video again and talk about which observable movements constitute the behavior. Through this discussion and a frame-by-frame analysis of the behavior, the class develops a precise operational definition of feeding and high interrater reliability can be demonstrated. This analysis is much simpler and more precise than would be possible by replaying a videotape. In addition, students find it amusing and engaging if the data entered before class makes the computer appear to inaccurately code the behavior.

Although these data do not allow us to determine the cause of the group difference on the laboratory reports, one explanation is that COR provided a combination of attention-grabbing graphics, precise control, and student input into the lessons that enhanced the COR-instructed students’ performance. Furthermore, the COR-instructed students were more positive about the importance of actual research experience. Perhaps students in the COR-instructed group had a clearer understanding of the precision and control required to do “good” science, thus making it appear more important. Or perhaps the engaging quality of the presentation created a general halo effect having to do with research experience.

The advantages seen in the COR-instructed class are consistent with other research showing that computer technology can enhance learning, especially if it goes beyond simple drill and practice text-based approaches (e.g., Forsyth & Archer, 1997). We should note, however, that we found no difference between the two groups on the multiple-choice exam. The similar performance on the exam likely reflects that both classes studied from the same reading material and covered the same material in class as well as that the questions on the test were primarily definitional. This finding is consistent with Welsh and Null (1991), who also found no difference on a comprehension task between students taught using computer simulations and those taught using traditional methods. These mixed results confirm the importance of looking at what the instructor expects students to get from the experience when assessing effectiveness of software (Castellan, 1993). If we had based assessment solely on exam performance, we would not have seen that students who used COR were better able to apply the concepts in their research and were more convinced of the value of research.

**References**


Notes

1. Funds for the development and modification of Courseware for Observational Research came from a Faculty Technology Initiative Grant from Pennsylvania State University's Educational Technology Center and development funds from Penn State Erie, The Behrend College.
2. We thank the Penn State Educational Technology design team for Courseware for Observational Research: Eric Corty, Carl Kallgren, Carol Dywer, Morris Weinstock, and Barbara Polka Smith; the students and faculty who gave us feedback that allowed us to improve the program; and especially Charisse Nixon for being the guest lecturer in this study.
3. Instructors interested in using Courseware for Observational Research should contact Dawn Blasko, Penn State Erie, The Behrend College, School of Humanities and Social Sciences, Station Road, Erie, PA 16563; e-mail: Dawnblasko@psu.edu.
4. Send correspondence to Victoria Kazmerski, Penn State Erie, The Behrend College, School of Humanities and Social Sciences, Station Road, Erie, PA 16563; e-mail: vak1@psu.edu.
Experiential activities help psychology students experience firsthand the richness and depth of meaning of qualitative data. Through these activities students grapple with pragmatic, philosophical, and ethical questions like those facing researchers in the field. These activities concern focus groups, observation, data collection, cultural sensitivity, ethnomethodology, data analysis, and ethics. In a final exercise students develop answers to the common challenges posed by skeptics of qualitative methodology. Instructors can select single exercises for broad psychology courses or use all of them, as we do, in a course dedicated to qualitative research in psychology.

Social scientists increasingly rely on qualitative data as both an adjunct to quantitative data and a valid form of information on its own (Patton, 1996). Some psychology departments teach a separate qualitative research course, and others include qualitative research orientations in more traditional research methods courses. We teach a one-semester graduate course on qualitative research to help students know when qualitative approaches are most useful, have the skills to conduct rigorous qualitative studies, and know how to evaluate qualitative research—none of which are as intuitive as they might seem. In this article, we describe exercises that we use in our courses to convey the nature and variety of qualitative research in psychology, in the hope that other teachers of qualitative research will find them useful. Given the large number of approaches used in qualitative research, these exercises cover selected techniques only and are not exhaustive.

Rossman and Wilson (1985) described the major ways of combining qualitative and quantitative data (words and numbers) as initiation, corroboration, and elaboration. Frequently, researchers begin their query by using qualitative methods to observe and describe the phenomenon of interest, which they then test empirically (initiation). Corroboration occurs when researchers bring together data collected through more than one method to determine whether there is convergence in the findings. Finally, one type of data can be used to elaborate on the findings of another, either lending strength to an argument or uncovering contradictions (elaboration). At the end of our course, many students express interest in conducting studies with both qualitative and quantitative components. Others decide that one approach only fits better with the problems that interest them.

When we teach qualitative methods, we try to highlight the advantages of qualitative research without constantly referring to its more established quantitative cousin. Different methods are valuable for different kinds of questions, and we hope our students will learn to judge study designs on their appropriateness for the problem—not on some a priori preference for one approach or another. Over the years, we have seen journals become more receptive to qualitative studies (Banyard & Miller, 1998), and we have encountered increasing numbers of students who hope to conduct at least one qualitative study—without necessarily connecting it to a quantitative study—during the course of their graduate education. For this reason, we have developed this graduate course dedicated exclusively to the methods and techniques of qualitative research. Our class meets for one 2½ hr session each week, which enables us to conduct at least one experiential exercise in each class. Instructors whose classes follow different schedules will have to adapt the exercises accordingly. Instructors may wish to use specific exercises described here in classes on research methods or other topics. For instance, teachers of health psychology courses can use the focus-group exercise to familiarize students with this common health psychology research technique. Similarly, teachers of community psychology courses may discover that the exercise exploring the cultural sensitivity of qualitative research fits well (Maton, Hrabowski, & Greif, 1998).

Setting the Stage

By the time they enter our graduate course in qualitative research, most students have been taught that their usual ways of knowing the world—through observation and conversation—are not suited to the academic study of psychology and that facts necessarily include numbers (Hubbard, 1988). They have learned that scientific knowledge may be obtained only through reductionist procedures and statistical manipulation and that rigor comes with large sample size, standardized validated instruments, and numerical measures of reliability and
validity. Some students attend our classes with the suspicion that qualitative research is somehow weaker (or easier) than quantitative research. Immersion in the challenging exercises described here relieves students of these notions.

Techniques that actively involve students facilitate learning (Older, 1979). We begin our courses in qualitative research with a series of demonstrations of the value of nonnumerical information to help counteract some of the “the truth is only in the numbers” messages that our students have received in prior psychology classes. For instance, we ask them how they chose their program of study (their comments inevitably include “impressions” and discussions with students and faculty but rarely include statistics). We ask them how employers choose among job candidates. Again, students remind us that interviews are usually paramount. Finally, we ask them if they would trust any standardized battery of instruments sufficiently to choose their life partner based solely on the results. By this point, students begin to see how much they rely everyday on primitive qualitative inquiry and usually express a greater openness to learning about how to make qualitative research more rigorous.

In the first class, we sometimes ask each student to rate a teacher, department, or job on a 5-point scale ranging from 1 (very poor) to 5 (excellent) and explain their answer to a classmate. Then, in the large group, we ask for reactions to the two types of data (rating and verbal explanation). Usually, the discussion highlights the depth, complexity, and contextual information that become available in descriptions as compared to the breadth afforded by numerical ratings. Also, at the beginning of the semester we ask class members to introduce themselves by telling their name and describing a turning point in their lives. We then discuss the power of this information and how qualitative research can capture that power and convey it directly to readers. In this way, even skeptical students begin to see the value of qualitative data.

Using Focus Groups to Introduce Focus Groups

Psychology researchers frequently use focus groups to determine community needs, obtain group interview data, and evaluate proposed research protocols and mental health services. We want our graduates to know how to conduct focus groups, either as a major source of data or as an adjunct to other methods. In this exercise, students gain experience conducting or participating in a focus group.

After a short introduction to the definition and history of focus groups (Krueger, 1994), we ask for enough volunteers for each to be a leader of a minifocus group of four or five class members. The purpose of the focus group, we tell them, is to explore components of effective teaching. We then briefly tell the volunteers about their role as focus group leaders (e.g., present ground rules, put the group at ease, encourage different opinions, bring the group back to the topic at hand). Our instructions take no more than 10 min. The volunteers then use a question guide that we give them (available from authors) to lead their minifocus groups. We stop the groups after about 25 min, although they have not yet discussed all the questions. We ask for personal reactions and answer questions that arose from the experience (e.g., “What do I do if someone doesn’t talk?” or “Is it OK if one group member asks another a question?”). We have found that this firsthand experience makes students more receptive to our minilectures that follow on the uses, strengths, weaknesses, and mechanics of focus groups (Krueger, 1994; Piercy & Nickerson, 1996). Students understand the advantages and limitations of focus group interviews and choose to conduct studies using focus groups with greater frequency after learning about them in this course.

Observing and Writing Field Notes

As part of their training in the social sciences, students learn to draw inferences. Although drawing inferences may be helpful at the data analysis stage, this is a habit to keep at bay while gathering data. Students need to learn to differentiate between their observations and inferences. For example, students trained in the rudiments of clinical psychology who see a couple sitting on opposite ends of the couch with their arms crossed are likely to conclude that there is marital conflict; if they see a child with bruises, they may suspect child abuse; and if they see someone who is frowning, with downcast, tearful eyes, they may conclude that this person is depressed. Although these inferences may be correct and may be useful in a clinical situation, they can interfere with the students’ ability to simply describe what they see. In this exercise, we encourage students to become sophisticated nonjudgmental observers without interpreting or drawing conclusions.

First, we have them compare the following two reports on the same person:

1. A rich man got onto the elevator. He looked like he was rushing to a job interview.
2. The first person to step into the elevator was a bearded brown-haired man who wore a dark blue suit, white shirt, and shiny shoes. He clutched a polished leather briefcase to his chest and looked at his watch twice while on the elevator. After pressing the button for the eighth floor, he examined his reflection on the metal elevator wall and straightened his tie. He shifted from foot to foot and squirted breath spray into his mouth.

Obviously, the writer of the second report described without inferring. This, we tell them, is what we want them to do in the observation exercise.

We give students 15 min in which they are to stand or sit in a place where they can see people (e.g., in a hallway, in front of an elevator, under a tree) and simply observe. We tell them to interact with people at the place as little as possible. No more than two students from the class can be at the same place. We ask them to write down what they observe, sticking as close to actual description as possible.

After they return to the classroom, we ask them to write field notes for 10 min about the experience. We instruct them to include both practical information (e.g., where they conducted the observation, for how long, their reactions to it) and their feelings about it (e.g., “I was nervous about what people would think about my just standing there”). We ask several
students to read the field notes about their experiences to the entire class. They usually describe the observation task as both much harder and much richer than they anticipated. We then have several volunteers read their observation descriptions aloud. We talk about what makes a description vivid and how to choose which details to record at the moment and which to save for the field notes. We also spend some time talking about the “self” issues that emerge in an observation (e.g., feeling awkward, feeling like joining in a conversation, feeling affinity for some people and enmity for others). After this exercise, student reports show greater attention to relevant details in their observations. They also demonstrate increased awareness of their use of self in research and an improved ability to describe without editorializing.

**Experiencing Various Methods of Data Collection**

In this class activity, students learn about different types of qualitative data and gain practice collecting data. We organize a focus group discussion on an engaging topic and then assign five groups to observe the discussion and collect five different types of qualitative data (focus-group discussion guide available from authors). We discuss how these data might be used in a qualitative research study.

First, we ask for five or six female volunteers who are willing to take part in a short (about 30 min) minifocus group on the subject of gender bias (another topic may be chosen, but we have found this one generates lively discussions). We then divide the rest of the class into five teams and arrange the minifocus group in the center as a fishbowl with the five observing teams on the outside. The first of the five teams observes the focus group discussion and inductively derives observer-constructed categories that emerge as the discussion proceeds (i.e., they are to take notes on the discussion and then group the topics that emerge in ways that seem to make sense). The second team observes and identifies the indigeneous typologies (Patton, 1990, p. 393) that emerge (e.g., the speakers themselves may describe three categories of sexist bosses as “the kind of boss who cares about your work,” “the kind of boss who can’t take your work seriously,” and “the kind of boss who’s always putting you on the defensive”). The third team writes field notes in which they describe the setting and maintain a general record of the interaction dynamics and the focus-group topics. The fourth team writes a reaction/feeling diary of their reactions as the focus group discussion unfolds. The fifth team writes metanotes, field notes focusing on the process of the learning activity itself.

After the focus-group discussion and a brief follow-up discussion by the members about their experience, each of the observing teams meets separately to discuss what they recorded individually. Each observing team picks a spokesperson who summarizes for the class their team’s experience and the specific data that their team collected. We then lead a discussion on the various types of data and how these might be used in an article on this focus-group study. Instructors may also want to take advantage of this opportunity to discuss how the researchers’ relevant identity groups (in this case, men or women) might affect the data-gathering and analysis process. After this exercise, students demonstrate improved ability to differentiate among the various categories of information researchers can gather during a study, which makes their research proposals more precise.

**Cultural Sensitivity of Qualitative Research**

As a field, psychology is increasingly concerned with developing methods that address humanity’s cultural diversity (Matsumoto, 1996). In this exercise, students see how qualitative methods are well suited for studying cultural issues because participants have opportunities to elaborate on the meanings behind their cultural practices. In qualitative studies, participants have an opportunity to establish their own categories and place emphases where they wish. People from all cultural groups habitually engage in the most common methods of data gathering used in qualitative research: observation and conversation. We share with students our observation—obtained through fieldwork in Latin America, Asia, and among immigrants in the United States as well as our reading of the crosscultural literature—that researchers often err when they assume that the same quantitative instrument can be translated and used crossculturally. Written instruments with predetermined categories, particularly when they are used alone, might miss the mark or seem foreign or alienating to people from different cultures (Fontes, 1997; Loos, 1995).

In this exercise, we ask the class how they might learn about affection toward children in an ethnic subculture in their community. They usually describe observing families and conducting interviews with their members, individually or collectively. We then discuss the advantages and disadvantages of following up on this initial fieldwork by asking open-ended qualitative questions (e.g., “How do you show your child you’re proud of him or her?”) or quantitative questions (e.g., “How many times a day do you hug your child?”). On a scale of 1–5, how important is it to hug your child? As long as they are implemented in a culturally sensitive manner, neither method is better; they simply provide different kinds of information. If there is sufficient time in the course, students implement this exercise with a small number of participants. This exercise leads students to incorporate attention to culture into their research.

**Ethnomethodology**

We want our students to understand the implicit cultural assumptions that shape their and others’ world views. In this way, we hope they will be able to conduct research that acknowledges and explores different cultural positions. The following ethnomethodology exercise provides a fine vehicle for this lesson. Ethnomethodologists study ordinary, routine, everyday behavior that is usually invisible because people take it for granted (e.g., ethnomethodologists might consider what a Martian would need to learn about flirting to function appropriately at a high school dance). To access unspoken norms, ethnomethodologists often conduct ethnomethodological experiments that disrupt normal routines by doing something out of the ordinary (Garfinkel, 1967). We ask students to choose one of the subsequent tasks or to propose one of their
own and write a short paper describing their reactions to conducting the out-of-the-ordinary experiment, identifying the tacit knowledge or rules that might have surfaced. Possible ethnomethodology experiments include the following:

1. Get into a crowded elevator and stand with your back to the door.
2. Challenge someone to a game of tic-tac-toe. When the other person marks an X or O, erase it and put it in another place, then proceed with your own mark.
3. Stand within 6 in. of someone you know slightly and begin a normal conversation.
4. Start singing in a public bus.
5. Introduce yourself and shake the hands of strangers in a line to buy movie tickets.
6. Talk to the fruits or vegetables in a supermarket.
7. Wear pajamas all day as you go about your regular activities.

Students should discuss all proposed experiments with the instructor or the entire class because students sometimes come up with ideas that are inappropriate or ethically questionable. These discussions provide fruitful ground for more general conversations about research ethics and acquaint students with the importance of peer review and prior approval.

As an example of an ethnomethodological investigation, one of our students got approval from us and the manager of a supermarket to sample various foodstuffs (e.g., drinking from a milk carton) prior to buying the items. The manager agreed as long as the student subsequently paid for the items he sampled. The student was surprised both at how hard it was for him to engage in this mildly taboo activity and also at the strength of the stares of reprobation that he encountered from other shoppers. We have found these exercises to be useful, engaging ways to sensitize our students to implicit rules and norms, leading them to conduct research that takes into consideration their own and their participants’ implicit cultural biases.

Data Analysis

As the major assignment in the course, students conduct and write up a small-scale qualitative study. Students often use this assignment as an opportunity to test an approach for their thesis or dissertation. A student who expects to interview 12 clinical psychologists and supervisors for a dissertation, for example, might interview 1 of each for this project.

To gain practice in data analysis, we have a student bring in two pages of an interview transcript from this project, disguising identifying information and leaving plenty of room in the right margins. (Students volunteer eagerly to bring in data because it gives them a head start in analyzing data for their project.) We distribute copies of the transcript to everyone in the class. We then ask the students to do a line-by-line reading of a page, individually listing all of the themes they can identify. After 10 min, we ask them to break into small groups and compare notes. As a class, we list all of the themes identified by each group. We have found that the themes identified often vary widely. We then ask the students to return to their annotated transcripts and process the full two pages this time, determining which themes could appropriately be grouped into more abstract or larger categories (e.g., anger, helplessness, and frustration could be grouped under the heading “feelings about the work”). We compare notes again as a class and discuss the integrity of the categories, what is lost and gained by collapsing subcategories, which subcategories should be split into still smaller categories, and so on.

This exercise accomplishes a number of objectives. First, it gives students the hands-on experience of analyzing a transcript. They are able to see how themes emerge from the data, and at the same time it becomes clear how readers’ perspectives affect the themes they identify. They also learn about the process of working with others on data analysis. They are able to see how many more themes emerge when several people work on the data and how much rich thinking can emerge from such discussions. Data analysis ceases to be a mystical process and is revealed to students for what it is: a laborious process involving hard work, keen attention, collaboration, persistence, and tolerance for ambiguity. This exercise alerts students to the various levels at which data can be coded and the need to read manuscripts numerous times. Recently, we have found that some students dedicate more time to learning complicated qualitative data analysis programs than to immersing themselves in the data, which we believe is a mistake. This in-class intensive immersion in two pages of data helps them see what they can mine manually. They learn that in the final analysis (figuratively and literally) they are the most important instrument in qualitative data analysis. Moreover, they learn to trust their ability to analyze data and are able to approach this phase of research with greater competence and confidence.

Morals and Ethics in Research

Reynolds (1979) distinguished between moral concerns in research, defined as “acting in accordance with accepted notions of right and wrong” (p. ix) and ethical concerns, which generally relate to conformity to a code or set of principles established by a professional organization. Moral principles are more personal and likely to stem from family, religious, and cultural values. Institutional review boards and the Ethical Principles of Psychologists (American Psychological Association, 1992) concern themselves with ethical decisions, which are often narrower and easier to resolve than the moral dilemmas involved in research. We want our students to grapple with both ethics and morals in research. Through the following exercise, we prepare our students to think deeply about moral and ethical issues, seek direction in professional guidelines and through discussion with colleagues and supervisors, and ultimately make their research decisions with as much ethical and moral sensitivity as possible.

Prior to this exercise, we assign reading materials on research ethics in psychology (e.g., Fontes, 1998; Keith-Spiegel & Koocher, 1998). In class, we present for discussion a series of 15 scenarios that represent ethical dilemmas (available from the authors). In the interest of time, we divide students into small groups to focus on two dilemmas each. After the groups have enough time to discuss their dilemmas as well as
what they would do and why, each small group presents the main point of their discussion to the entire class. Often students comment that the ethical guidelines that they have read do not apply easily to the qualitative research dilemmas they are assigned. Three of the dilemmas we use for this exercise follow:

1. You are volunteering or working in a nonresearch setting (e.g., at a rape crisis center or at a restaurant) and in the course of your work decide that this would be a good place to collect qualitative data through informal interviews and observations. You do not want to tell people about your study because you would be speaking with them and observing them anyway. What are the ethical implications of this situation? (see Fine, 1992)

2. You are meeting with scholars in another country where you are planning to conduct research interviews. The experts from that country advocate hiding the tape recorder during the interviews to avoid making the participants nervous. There is no Human Subjects Review Board in that country. What do you do?

3. You are evaluating an educational program for preschool children, a program that you think is useful, if not perfect. In the process of conducting your study you learn that the directors of the program are falsifying rates of completion to maintain their levels of funding. If this information were to become public, the program would be forced to shut down. What do you do with this “guilty knowledge?” (Gottfried, 1996)

After completing this exercise, students report that questions of ethics and morals are more complicated than they had realized. They discover that review boards sometimes approve projects that fall within ethical guidelines but are morally questionable, different cultural groups often hold different moral views of a project (Fontes, 1998), and the moral correctness of a project is often disputed by the various stakeholders.

Defending Qualitative Research

We hope that by the end of the course the students will be able to provide a credible rationale for using qualitative research in psychology. This skill may help them handle the challenges qualitative researchers sometimes confront in their thesis and dissertation committees, grant proposals, and submissions of articles for publication in psychology journals. Toward the end of the course, we require our students to read several articles that outline strategies for defending qualitative methods (e.g., Dreher, 1994; Marshall & Rossman, 1989, chap. 5). In the next class, we hand the students a brief description of a qualitative study in psychology. We like to use a paragraph about a study that one of the students in the class is proposing, with that student’s permission. (We tell the students they can invent information about the study that may be missing from the brief description.) We divide the class into small groups and ask each group to prepare responses to one of the categories of the subsequent questions. The task for these small groups, then, is to develop answers to their assigned questions that adequately and realistically defend qualitative methodology in general and this study in particular. Some of their answers may involve debunking common misperceptions about qualitative research, whereas others involve ensuring the rigor of the specific proposed study. The readings and discussions of the entire course provide them with the information they need (we cite particularly relevant articles next to each question). When we come back together as a large group, students (in the spirit of fun) play the role of skeptical adversaries, asking each group, in turn, the questions they have been assigned. The entire class then discusses or elaborates on their answer before the next question. The questions we assign follow.

1. Post-modern perspective: How can you say anything definitive about your qualitative research findings? Much qualitative research relies on the postmodern perspective, which contends that “reality is in the eye of the beholder.” If everything goes and no reality is privileged, how can you be sure of anything? What’s the point? (Gergen, 1988)

2. Generalizability and sample size: If qualitative research can’t be generalized, what good is it? How can we have any faith in findings that can’t be generalized? Wouldn’t it be better to consider this study a pilot and to conduct a large scale survey afterward? With such a small sample, how can this study be useful? (Kennedy, 1979)

3. Researcher bias: What the researcher attends to, the questions the researcher asks, even the quotes the researcher chooses to put into a report could all bias that report. So, how can we trust a qualitative research report? What can you do to reduce researcher bias? Isn’t it true that qualitative studies will always be able to produce exactly what the investigator expects? (Lather, 1986; Phillips, 1990)

4. Rigor: Social scientists need solid quantitative proof to gain equal footing with physical scientists. Good quantitative instruments have been validated. Qualitative research is just opinion. Shouldn’t we seek objective, factual data instead of subjective opinions? What is the difference between qualitative research and journalism? (Reason, 1994)

5. Reliability and validity: Is this study replicable? If not, how can it be considered scientific? How do you address reliability and validity in qualitative research? (Kirk & Miller, 1986)

6. Subjectivity of the participants: If you use qualitative evaluation methods, how do you know whether a program is truly effective? Maybe a program is good for the people who use it, but they may not like it. If you use qualitative methods you might get an overly negative view of this effective program. Don’t you agree? (Maguire, 1987)

7. Legitimacy: Real science uses statistics. We need to teach our students how to conduct legitimate scientific experiments. What is the point in allowing you to proceed with this study, which would never get funded or get published anywhere? Isn’t it true that qualitative methods do not have legitimacy or acceptance today? Don’t you want tenure? What would you do to increase
the chances that this qualitative proposal will be funded and that an article about the research will be accepted for publication? (Denzin & Lincoln, 1994)

Providing the answers to these questions is beyond the scope of this article. However, after a full semester’s study of qualitative research, our students approach these questions with authority and relish. They show pride in their ability to debunk common misperceptions about qualitative research, develop rigorous study designs, and explain their design decisions with convincing rationale.

Conclusions

A number of indicators point to the effectiveness of this class. First, students tend to rate the course positively, giving it high marks both for usefulness and interest. In their written evaluations, they make comments affirming the utility of the class, such as “You taught me that interviewing well is a lot more complex than just asking and listening.” Second, our colleagues have affirmed the importance of the course by refusing to chair qualitative dissertations unless the student has taken a graduate-level course in qualitative research. Finally, graduates of the class often choose to write rigorous theses and dissertations based solely or partly on qualitative data. Articles based on these studies have been accepted into refereed journals.

Most of the activities in this article reflect our belief that people learn best through direct experience. It is not enough to explain that qualitative data are rich and culturally sensitive—students need to experience that richness and sensitivity. It is not enough to explain qualitative procedures, present the rationale for qualitative research, and discuss ethics. The instructor should also provide challenging activities that encourage students to use qualitative methods, reflect on them, and explore complex dilemmas. We hope that the activities presented here will be helpful to other teachers of qualitative research in psychology. We also hope they will stimulate creative instructors to develop additional ways to immerse psychology students in the fascinating, multi-faceted process of qualitative inquiry.

References


Note

Send correspondence to Lisa Aronson Fontes, School of Education, SDPPS, University of Massachusetts, Amherst, MA 01003; e-mail: LFontes@javanet.com.
Teaching Students about Classic Findings on the Detection of Deception
Kathryn A. Morris
Teaching of Psychology 2003 30: 111
DOI: 10.1207/S15328023TOP3002_05

The online version of this article can be found at:
http://top.sagepub.com/content/30/2/111
METHODS AND TECHNIQUES

Teaching Students About Classic Findings on the Detection of Deception

Kathryn A. Morris
Butler University

I describe a classroom exercise that demonstrates people's inability to detect deception better than chance. In the exercise, students worked in pairs and took turns asking each other a series of questions. Students lied to their partners some of the time, and they in turn attempted to determine when their partners lied to them. Students also reported their degree of confidence in their lie detection abilities. The results consistently replicated classic findings in this area of research. In addition, students who participated in this exercise retained the knowledge gained from their participation for 6 to 30 months. I discuss the courses in which this exercise might be useful.

Social psychologists often teach students about a wide range of optimistic biases that lead people to hold overly positive views of themselves and the world (e.g., Kunda, 1987; Weinstein, 1989). One specific bias covered in many social psychology textbooks (e.g., Aronson, Wilson, & Akert, 1999; Brehm, Kassin, & Fein, 1999; Myers, 1999) is that people tend to believe they are better at detecting others' lies than they actually are.

Although a few individuals may detect lies significantly better than chance, most people cannot (Ekman & Friesen, 1975; Ekman & O'Sullivan, 1991). Despite these findings, people remain overly optimistic about their lie detection abilities: People think they are quite good at detecting others' lies (DePaulo, Charlton, Cooper, Lindsay, & Muhlenbruck, 1997; Swann & Gill, 1997). In addition, people believe that the more confident they are in their lie detection abilities, the better able they are to detect lies, yet no such relation between confidence and accuracy in lie detection exists (DePaulo et al., 1997).

When social psychology instructors attempt to teach students about these classic findings in the lie detection literature, they face the additional problem that college students hold greater optimistic biases than people of other ages (Quadrel, Fischhoff, & Davis, 1993). Thus, although students may be able to learn that others are not very good at detecting lies and that confidence and accuracy in lie detection are not related, they may continue to believe that they are exceptions to these general findings. Therefore, it was my goal to create an exercise that would replicate classic findings in the lie detection literature and, as a result, teach students three principles: (a) that people cannot detect others' lies better than chance, (b) that confidence and accuracy in lie detection are uncorrelated, and (c) that the students, themselves, cannot detect others' lies better than chance.
search on lie detection (Ekman & Friesen, 1975; Ekman & O’Sullivan, 1991), providing additional evidence that although the majority of people cannot detect lies better than chance, a few individuals may be able to do so.

Next, to determine whether each class, as a group, detected lies better than chance, I compared the average number of correct judgments to four (four out of eight judgments should be correct by chance alone) using a t-test. In the semesters in which I conducted this exercise, class accuracy in detecting lies (M = 2.56, SD = 1.41; M = 2.79, SD = 1.14; M = 4.42; SD = 1.59; M = 4.58, SD = 1.42) never exceeded chance (ps > .05). To determine whether each class was overly confident in its ability to detect lies, I compared the average confidence score to 50% (the likelihood of correctly guessing each answer) using a t-test. Class levels of confidence (M = 61%, SD = 9.74; M = 63%, SD = 19.23; M = 64%; SD = 12.66; M = 68%, SD = 15.41) were always significantly greater than 50% (ps < .05). To determine whether confidence and accuracy in lie detection were related, I correlated average confidence level with average number of correct judgments, and these correlations never reached statistical significance, r(26) = .15; r(26) = .24; r(16) = -.25; r(14) = .34 (all ns). I conducted all statistical tests prior to the next class period.

Presenting the results to students. I spent 20 min of the next class period explaining that the results of the exercise mirrored classic research on lie detection. Despite the performance of a few individuals, if any, who detected lies better than chance, the class as a whole did not. Also, because the class’s average confidence level was statistically higher than 50%, I explained that students were overly confident in their ability to detect deception and that their confidence in their lie detection abilities was not correlated with their accuracy on the lie detection task. I also emphasized that the exercise was biased in favor of the students: They were familiar with their partners, and they knew their partners would lie to them. Yet, despite these benefits, the class could not detect lies better than simply guessing.

Discussion

Although this exercise consistently produced data that replicated classic findings on lie detection, it is important to determine whether students learned and retained the principles exemplified by the results of the exercise: (a) that people cannot detect others’ lies better than chance, (b) that confidence and accuracy in lie detection are uncorrelated, and (c) that the students, themselves, cannot detect others’ lies better than chance. To evaluate whether students learned and retained these principles, I compared a subset of students who participated in the exercise with students who had not participated in the exercise. Both groups of students completed a questionnaire designed to assess their knowledge of classic findings on lie detection.

Evaluation of Exercise

Method

Participants. Thirty-two psychology students participated in the evaluation of the exercise. Fifteen of these students (3 men, 12 women) never participated in the exercise. The remaining 17 (1 man, 16 women) participated in the exercise. Of these 17 students, 8 participated in the exercise 6 months earlier; 4 participated 18 months earlier, 3 participated 24 months earlier, and 2 participated 30 months earlier (M time lapse = 14.8 months; SD = 9.0 months).

Procedure. I approached students in the psychology department student lounge and asked them to complete a questionnaire to assess student learning as a result of a classroom exercise on lie detection I had implemented in Social Psychology. On the questionnaire, students first indicated whether they participated in the exercise. They then reported the percentage of the time (0% to 100%) they thought most college students could accurately detect others’ lies. Next, they answered an open-ended question about the relation between confidence and accuracy in lie detection. Finally, students reported the percentage of the time (0% to 100%) they thought they could accurately detect others’ lies. Students who participated in the exercise also reported on scales ranging from 1 (not at all) to 7 (very) how enjoyable and memorable they found the exercise to be.

Results

First, students reported the percentage of the time they thought most college students could accurately detect others’

Table 1. Questions Used in the Classroom Exercise

<table>
<thead>
<tr>
<th>List 1</th>
<th>List 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. What is your favorite type of pizza?</td>
<td>1. What is your favorite dessert?</td>
</tr>
<tr>
<td>2. Of all the television shows you’ve seen lately, what is your favorite?</td>
<td>2. Of all the movies you’ve seen lately, what is your favorite?</td>
</tr>
<tr>
<td>4. What activities did you participate in as a high school student?</td>
<td>4. What extracurricular activities are you involved in here at (name of college)?</td>
</tr>
<tr>
<td>5. What is your favorite course thus far at (name of college)?</td>
<td>5. What was your favorite class in high school?</td>
</tr>
<tr>
<td>6. What do you hope to do after graduation?</td>
<td>6. What sport do you play the best?</td>
</tr>
<tr>
<td>7. What animal do you think is the ideal pet?</td>
<td>7. When you go to the zoo, what animals do you like to see the best?</td>
</tr>
<tr>
<td>8. Which academic subject is your best?</td>
<td>8. What sport do you play the best?</td>
</tr>
</tbody>
</table>

Note. All participants received both lists. One participant in each pair prepared answers for the questions on List 1 and asked his or her partner the questions on List 2; the other participant in each pair prepared answers for the questions on List 2 and asked his or her partner the questions on List 1.
lies. Students who had not participated in the exercise believed college students could detect lies a greater percentage of the time ($M = 61\%, SD = 11.6$) than students who participated in the exercise ($M = 41\%, SD = 18.9$), $t(30) = 3.55, p < .01$. Furthermore, among students who had not participated in the exercise, this estimate of $61\%$ accuracy in lie detection abilities was significantly greater than $50\%$ accuracy, which would be expected by chance alone, $t(14) = 3.57, p < .01$. However, among students who participated in the exercise, this estimate of $41\%$ accuracy was marginally less than $50\%$ accuracy, which would be expected by chance alone, $t(16) = -2.04, p < .10$. Clearly, students who participated in the exercise learned that people cannot detect lies better than chance (and may actually be worse than chance).

Next, students answered an open-ended question about the relation between confidence and accuracy in lie detection abilities. After grading answers on a correct–incorrect dichotomy, a chi-square analysis revealed that significantly more students who participated in the exercise (71%) correctly understood the (non)relation between confidence and accuracy in lie detection ability than did students who had not participated in the exercise (13%), $\chi^2(1, N = 32) = 10.61, p < .01$.

Students then reported the percentage of the time they thought they could accurately detect others’ lies. Students who had not participated in the exercise believed they could detect lies a greater percentage of the time ($M = 67\%, SD = 15.3$) than students who participated in the exercise ($M = 46\%, SD = 21.5$), $t(30) = 3.20, p < .01$. Furthermore, among students who had not participated in the exercise, this estimate of $67\%$ accuracy in lie detection abilities was significantly greater than $50\%$ accuracy, which would be expected by chance alone, $t(14) = 4.39, p < .01$. However, among students who participated in the exercise, this estimate of $46\%$ accuracy was statistically equivalent to $50\%$ accuracy, which would be expected by chance alone, $t(16) = .79, ns$. Clearly, those who participated in the exercise learned that they, themselves, could not detect lies better than chance.

Finally, students who participated in the exercise rated it as being both enjoyable ($M = 5.88, SD = 0.81$) and memorable ($M = 5.38, SD = 1.20$).

**General Discussion**

In addition to producing data that replicated classic findings on the detection of deception, this exercise clearly accomplished its three learning objectives. As a result of participating in the exercise, students learned that people are generally unable to detect others’ lies, that confidence and accuracy in lie detection are uncorrelated, and that they, themselves, were unable to detect lies better than chance. In fact, student retention of these principles ranged from 6 to 30 months, lasting an average of 14.8 months. This exercise is effective, easy to use, and enjoyable to students.

In my social psychology courses, I use this exercise to lead into a general discussion of biases in the person perception process and past research on this topic. For example, I explain that although the lies used in this exercise are consequential, studies of more consequential lies (e.g., lies about sexual behaviors) yield similar results (Swann & Gill, 1997). This exercise could also be used when teaching about nonverbal behavior. For example, when attempting to detect their partners’ lies, teachers could instruct students to pay attention to their partners’ words or to their nonverbal behaviors (see Fein & Spencer, 1999) and could analyze the resulting data to compare these two groups.

Although I designed this exercise for use in social psychology courses, it would also be useful in other courses. For example, because lie detection abilities are particularly relevant to law enforcement and judicial officers, this exercise may be a valuable addition to courses on psychology and the law. This exercise would also make a nice addition to research methodology and statistics courses. In such courses, students could serve as researchers who conduct the exercise as an experiment and analyze the resulting data. Doing so would reinforce classroom lessons on hypothesis testing and provide experience conducting simple statistical analyses.

**References**


**Notes**

1. Portions of this article were presented at the sixth annual American Psychological Society Institute on the Teaching of Psychology, Denver, CO, 1999.

2. I thank Ryan Brown, Brian Giesler, Charles Morris, Bob Padgett, the editor, and three anonymous reviewers for their helpful comments. In addition, I thank the students who participated in this classroom exercise.

3. Send correspondence and requests for copies of the materials used in this classroom exercise to Kathryn A. Morris, Department of Psychology, Butler University, 4600 Sunset Avenue, Indianapolis, IN 46208; e-mail: kmorris@emailbutler.edu.
Demonstrating the Importance of Question Wording on Surveys
Laura Madson
*Teaching of Psychology* 2005 32: 40
DOI: 10.1207/s15328023top3201_9

The online version of this article can be found at: http://top.sagepub.com/content/32/1/40

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Society for the Teaching of Psychology

Additional services and information for *Teaching of Psychology* can be found at:

Email Alerts: http://top.sagepub.com/cgi/alerts

Subscriptions: http://top.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Jan 1, 2005

What is This?
This classroom activity demonstrates the importance of question wording on surveys. In the activity, students compared data from 2 versions of a short questionnaire that differed only in the wording of the questions and the response scales. Class discussion focused on how wording affected the data and, consequently, how difficult it would be to draw valid conclusions from the data. After participating in the activity, students were better able to identify and correct poorly worded items and response scales on a subsequent survey.

Poorly worded questions or response scales can make it difficult or impossible to interpret self-report data (Schwarz, 1999). Small changes to the wording of questions or response scales can also significantly change the results or conclusions of a study (e.g., Catania et al., 1996; Wright, Gaskell, & O'Muircheartaigh, 1997; see also Wiederman, 2002). Given the frequency with which students encounter surveys (e.g., in the psychological literature, in the popular press), undergraduate psychology students should understand how question wording can affect self-report data.

What options do instructors have to teach about the importance of question wording? The teaching literature offers several techniques that illustrate various aspects of survey construction and methodology (e.g., Carducci, 1996; Gordon, 1987; Kerber, 1983; Millard, 1983; Thompson, 1994; Timmerman, 2000; Yoder, 1979), but none of these techniques directly demonstrates the effects of wording per se. This article describes an activity that shows students how the wording of questions and response scale can affect survey data.

**Method**

**Participants**

Participants (N = 23) were students in my sexual behavior class. They collected data from five friends using one of two versions of an eight-item survey. The two versions of the survey differed only in the wording of the items or the response scales (see Table 1). To bolster the sample size, I also collected data on 85 introductory psychology students, resulting in a total sample of 200 (104 participants completed Version 1; 96 completed Version 2). I collected all data anonymously, and all participants completed an informed consent statement specifying the voluntary nature of participation prior to taking the survey.

**Procedure**

I summarized the data by calculating mean responses for each item separately for the two versions of the survey. In a subsequent class period, we compared the results from the two versions and discussed how the data were affected by the wording of the questions or the response scales. The activity occurred in the second week of class, in conjunction with the chapter on research methods in the textbook. Students participated in the activity prior to any discussion of research methods in class. Instructors could opt to reverse this order by conducting the activity as a means of illustrating the key ideas in an earlier lecture on research methods in general or question wording specifically.

Class discussion focused on four issues: (a) Will participants agree on the meaning of a term? (b) Does the item make implicit assumptions about participants? (c) Does each question measure the concept of interest? (d) Will you be able to accurately interpret the data? Regarding the first issue, we discussed that the term sexual freedom had a number of possible meanings including the freedom to be sexually active, to choose one's sexual partners, to choose the timing and setting of sexual activity, or the freedom to be sexually active with more than one person. Similarly, we discussed different meanings of protection, what it means to have sex, different meanings of committed relationship, and how these different interpretations could influence participants' responses to the items.

Items 3 and 7 made implicit assumptions about participants, namely that they owned a car and were sexually active. We discussed how participants for whom these assumptions were incorrect might respond to each item and how those responses could make it difficult to accurately interpret the data. For example, participants who did not have a car might have responded that they strongly disagreed with the statements regarding auto maintenance or they might have responded to the item based on their best auto maintenance intentions. Other participants may have skipped the item entirely. Any of these possibilities would make it more difficult to ascertain the sample's general opinion regarding the behavior in question.

The activity also demonstrated a more conceptual problem related to wording: Does the item measure the intended concept? For example, Item 7 asked participants to indicate their level of agreement with the statement, “My part...
ner I and always use protection” or “Although I know it is important, sometimes I don’t practice safe sex.” We discussed that this item was probably intended to measure the frequency of condom use rather than agreement with either general statement. Similarly, Item 4 should have asked participants to indicate how often they lie rather than measuring agreement with statements about lying.

Item 8 illustrated the effects of imprecise question wording and different response scales. One version of this item asked participants to indicate how often they exercised using a 3-point response scale, with 17% of participants indicating they exercised infrequently, 48% indicating they exercised occasionally, and 35% indicating they exercised often. The other version used more precise wording in the stem (i.e., “In the last 6 months, how often have you engaged in at least 20 min of aerobic activity?”) and provided a 7-point response scale (e.g., almost never, less than once per week, once per week). We discussed that the less specific version left both the definition of exercise and the intended time frame open to interpretation by participants. Also, the 3-point response scale yielded data where almost half the sample reported exercising occasionally. On the other hand, we also discussed that the data from the more specific version of the item might also be difficult to interpret meaningfully because virtually equal numbers of participants responded using each of the 7 points on the response scale (e.g., 17% responded almost never, 13% responded less than once per week, 12% responded once per week, 15% responded twice per week).

Finally, we discussed how different interpretations of terms, items that make implicit or explicit assumptions, and imprecise wording of questions or response scales could make it difficult or impossible to draw valid conclusions from data. For example, we discussed response biases, noting that participants generally agreed with most of the statements, even those that were mutually exclusive (e.g., “I make it a practice to never lie” and “Like all human beings, I occasionally tell a white lie”; $M = 2.65$ and 2.37, $t(198) = 1.39$, ns. Items with compound statements (e.g., “Although I know it is important, sometimes I don’t practice safe sex”) also posed interpretive challenges because there was no way to know with which clause participants were agreeing (i.e., the importance of safe sex or their practice of safe sex). We also discussed the problem with using general statements like those in Items 3 and 7 because these items introduce unnecessary issues with social desirability. Well-intentioned researchers who used either question alone could easily draw inaccurate conclusions regarding participants’ priorities; the flaws in the question were readily apparent only when we compared the data from the two versions directly.

### Table 1. Items and Mean Responses From the Two Versions of the Activity

<table>
<thead>
<tr>
<th>Item</th>
<th>Version 1</th>
<th>Version 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I oppose raising taxes.</td>
<td>2.88</td>
<td>2.18</td>
</tr>
<tr>
<td>2. The primary task of the government should be to keep citizens safe from terrorism and crime.</td>
<td>3.19</td>
<td>2.50</td>
</tr>
<tr>
<td>2. The primary task of the government should be to preserve citizens’ rights and civil liberties.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. I regularly perform routine maintenance on my car.</td>
<td>2.90</td>
<td></td>
</tr>
<tr>
<td>3. Sometimes I don’t change the oil in my car on time.</td>
<td>3.27</td>
<td></td>
</tr>
<tr>
<td>4. I make it a practice to never lie.</td>
<td>2.65</td>
<td>2.37</td>
</tr>
<tr>
<td>5. Monogamy is important to me.</td>
<td>1.98</td>
<td></td>
</tr>
<tr>
<td>6. People should wait to have sex until they are in a committed relationship.</td>
<td>3.03</td>
<td></td>
</tr>
<tr>
<td>6. Sex can strengthen a new relationship.</td>
<td>4.60</td>
<td></td>
</tr>
<tr>
<td>7. My partner and I always use protection.</td>
<td>2.92</td>
<td></td>
</tr>
<tr>
<td>7. Although I know it is important, sometimes I don’t practice safe sex.</td>
<td>5.22</td>
<td></td>
</tr>
<tr>
<td>8. How often do you exercise?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. In the last six months, how often have you engaged in at least 20 min of aerobic activity?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Almost never</td>
<td>17%</td>
<td></td>
</tr>
<tr>
<td>2. Less than once/week</td>
<td>13%</td>
<td></td>
</tr>
<tr>
<td>3. Once/week</td>
<td>12%</td>
<td></td>
</tr>
<tr>
<td>4. 2 times/week</td>
<td>15%</td>
<td></td>
</tr>
<tr>
<td>5. 3 times/week</td>
<td>15%</td>
<td></td>
</tr>
<tr>
<td>6. 4 times/week</td>
<td>15%</td>
<td></td>
</tr>
<tr>
<td>7. More than 4 times/week</td>
<td>13%</td>
<td></td>
</tr>
</tbody>
</table>

Note. Version 1 is shown in standard text; version 2 is shown in italics. Unless otherwise specified, participants responded to each item on a 7-point, Likert-type scale ranging from 1 (strongly agree) to 7 (strongly disagree).
Learning Evaluation

Learning evaluation materials and procedure. The learning evaluation was a pre–post measure that students completed on the third day of class and a few days after participating in the activity \((n = 23)\). The measure consisted of 11 poorly worded items or response scales that might appear on a survey of sexual behavior (e.g., “Are you a virgin?” with a yes–no response scale and “How often do you have sex?” with a response scale ranging from less than once a month to daily). This instrument was different from the activity itself to demonstrate that student learning generalized to items related to course content. For each question, students described any problems with the question or response scale and suggested an alternative that would correct the problem.

Learning evaluation results. I matched students’ pretest and posttest responses using their self-reported birth date and favorite food. Three independent raters who were blind to condition coded students’ responses according to their overall quality. The raters coded each of the 11 questions separately resulting in 253 cases (i.e., 11 questions \(\times\) 23 students), each coded as either a success or a failure. Successes were those cases in which all three raters identified the posttest responses as higher in quality than the pretest responses. When the raters did not agree, the decision of two of the raters determined the outcome (e.g., if two raters identified a pretest response as higher in quality than a posttest response, the case was a failure). Responses for which students wrote the same remarks on both the pretest and posttest \((n = 17)\) cases were also failures (even if they were correct on the pretest). This analysis did not include blank responses \((n = 55)\). One hundred seventy-four out of the remaining 198 cases were successful, indicating that students’ descriptions of the importance of question wording were significantly better after participating in the activity (binomial test: \(p < .001\)).

Student Perceptions

Eight closed-ended items assessed students’ perceptions of the activity (see Table 2). Students’ perceptions of the activity were quite positive. They thought the activity was useful and recommended that I use the exercise again.

Students also described what they learned from participating in the exercise and what they thought was the best part of participating. Some of these comments included, “I learned how every word and the ordering of questions can affect survey results,” “the importance of operationalizing terms was made very clear,” and “that changing the wording of a question slightly can change people’s answers greatly.” A number of students commented on the difficulty of writing good survey items, noting that it is tough to describe the simplest thing and that the simplest things can turn out to be so complex.” Several students commented that the best aspect of participating in the exercise was comparing the two versions of the questionnaire, “directly observing the importance of question wording in surveys.” One student succinctly summarized the objective of the activity, noting that “wording makes all the difference.”

Discussion

Comparing the data from two poorly worded versions of a questionnaire helps students understand why the wording of questions and response scales is important. Students leave class with a better understanding of how poorly worded questions and response scales can bias data or lead to invalid conclusions. They also gain an appreciation for how difficult it can be to write good survey questions.

The activity has several advantages. Although I originally designed the activity for use in my sexual behavior class, the activity is appropriate for use in any class that discusses issues associated with self-report data or surveys (e.g., human sexuality, social psychology, personality psychology, industrial/organizational psychology, counseling psychology). Because it allows direct observation of the effects of wording on data, the activity is also simple enough for use in introductory psychology courses. Instructors can augment the in-class activity with homework that requires students to revise the poorly worded items to correct the flaws identified in class.

Instructors should be aware that this activity, like most, also has disadvantages. Although the in-class discussion requires only one class period, collecting and analyzing the data

<table>
<thead>
<tr>
<th>Item</th>
<th>Response Scale</th>
<th>Mdn</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>How much did you learn about the importance of question wording as a result of this exercise?</td>
<td>1 (very little) to 7 (a great deal)</td>
<td>7.0</td>
<td>0.79</td>
</tr>
<tr>
<td>How enjoyable was this exercise?</td>
<td>1 (not at all enjoyable) to 7 (very enjoyable)</td>
<td>5.0</td>
<td>1.30</td>
</tr>
<tr>
<td>The instructor should use this exercise again in future semesters.</td>
<td>1 (strongly disagree) to 7 (strongly agree)</td>
<td>6.0</td>
<td>1.09</td>
</tr>
<tr>
<td>How useful was this exercise in demonstrating the importance of question wording?</td>
<td>1 (not at all useful) to 7 (very useful)</td>
<td>7.0</td>
<td>0.85</td>
</tr>
<tr>
<td>I learned more about the importance of question wording today than if I had not participated in the exercise.</td>
<td>1 (strongly disagree) to 7 (strongly agree)</td>
<td>7.0</td>
<td>1.23</td>
</tr>
<tr>
<td>Today’s exercise was a waste of time.</td>
<td>1 (strongly disagree) to 7 (strongly agree)</td>
<td>2.0</td>
<td>1.11</td>
</tr>
<tr>
<td>Instructors who teach other, related courses should use this exercise when they discuss research methods.</td>
<td>1 (strongly disagree) to 7 (strongly agree)</td>
<td>6.0</td>
<td>1.24</td>
</tr>
<tr>
<td>Participating in the exercise increased my understanding of the importance of question wording.</td>
<td>1 (strongly disagree) to 7 (strongly agree)</td>
<td>6.0</td>
<td>0.81</td>
</tr>
</tbody>
</table>

Note. Medians are reported because the distribution of responses was negatively skewed.
does require some advance planning, particularly if instructors want students to collect the data themselves. Instructors could choose to collect the data themselves every semester or recycle a “stock” set of data across several semesters, although that approach may reduce students’ personal investment in the discussion. Conducting the in-class discussion in very large (N > 100) enrollment courses may be difficult. Students’ performance on the posttest learning assessment may also be a result of practice effects or learning gained from the textbook, rather than a direct result of participating in the activity.

Given their prevalence, all students will encounter a survey at some point in their lives. An understanding of the implications of poor survey construction may be the most practical lesson undergraduate psychology students can learn. This activity is a fun and effective way to teach students why wording makes all the difference.

References


Notes

1. Thanks to Dominic Simon, Adrienne Lee, Tara Gray, Rolston St. Hillaire, Dan Tappan, Kristina Long, Tunbi Adeogba, Joyce Hill, and three anonymous reviewers for suggestions that improved this article.

2. Send correspondence to Laura Madson, Department of Psychology, Box 30001/MSC 3452, New Mexico State University, Las Cruces, NM 88003; e-mail: lmadson@nmsu.edu.
Teaching Random Assignment: A Classroom Demonstration Using a Deck of Playing Cards
Craig K. Enders, Rick Stuetzle and Jean-Philippe Laurenceau
Teaching of Psychology 2006 33: 239
DOI: 10.1207/s15328023top3304_5

The online version of this article can be found at:
http://top.sagepub.com/content/33/4/239
Despite its apparent simplicity, random assignment is an abstract concept for many students. This article describes a classroom demonstration that uses a standard deck of playing cards. In small groups, students randomly assign playing cards (i.e., participants) to 2 treatment groups. Following randomization, students compare the relative frequency of "background variables" across the 2 groups (e.g., the number of red, black, face cards, spades). A pretest–posttest design indicated that quiz scores increased following the demonstration. We suggest a number of possible extensions for the demonstration.

Although seemingly simple, random assignment appears to be one of the more abstract concepts that students encounter when learning research design. In our introductory research methods courses, students write critiques of published research articles. In reading these critiques, we discovered that students understand how a lack of equivalence among treatment groups (i.e., selection; see Campbell & Stanley, 1963) can threaten internal validity, as they are quick to point out that preexisting background variables might influence the dependent variable. However, we also...
found that students do not understand the role of randomization in controlling for this threat, as they consistently cited group equivalence as a potential problem, even for true experimental studies that employed random assignment.

The apparent confusion over the value of randomization in experimental studies is not generally ameliorated by research methods texts. Textbook explanations of random assignment frequently appeal to the laws of probability in making a case for group equivalence. For example, Smith and Glass (1987) stated that, "The procedure results in an equitable distribution of subjects between the two groups because the series of coin flips constitutes a series of independent events, governed by the laws of probability" (p. 138). Christensen (2001) also referred to probability, stating that, "In order to provide equiprobability of events when randomly assigning participants to treatment conditions, it is necessary to use a randomization procedure" (p. 199). Similarly, Cook and Campbell (1979) explained that, "The equivalence achieved by random assignment is probabilistic" (p. 341). Other textbooks define random assignment or provide a description of the procedures used to implement randomization, but offer little explanation for how equivalence is actually achieved. For example, Gall, Gall, and Borg (2003) defined random assignment as "each sampling unit ... has an equal chance of being in each experimental condition," and subsequently concluded that, "Random assignment is the best technique available for assuring initial equivalence between different treatment groups" (p. 384). In our opinion, such explanations fall short because a clear understanding of randomization is predicated on knowledge of probability, which itself is an abstract concept to many introductory research students. We believe that a classroom demonstration may offer a more concrete introduction to the topic and therefore may enhance student learning.

A review of the literature revealed two previous papers that outlined random assignment teaching demonstrations. Labov and Firmaige (1994) had students work with a computer program that randomly assigned simulated participants to a specified number of treatment conditions. The researchers executed the simulation repeatedly and accumulated the results of the assignment procedure across replications. The goal of the simulation was to demonstrate that random assignment would result in equal frequency of assignment to the treatment conditions for every participant. Rothenberg and Sawilowsky (1999) used a Monte Carlo simulation to draw repeated samples of n = 4 from a large data array. They randomly assigned four observations to one of two groups, and they conducted independent t tests to assess differences among groups on simulated background variables. The goal of the simulation was to demonstrate the effectiveness of random assignment, as evidenced by the small proportion of statistically significant t tests across repeated trials.

Although a computer can illustrate the long-run behavior of random assignment, it largely obscures the underlying mechanism of randomization: Participants with differing characteristics are equally spread among the treatment groups such that the impact of confounding variables is approximately equal across treatments. Furthermore, the Rothenberg and Sawilowsky (1999) demonstration requires foundational knowledge of significance testing and statistical errors; introductory research students may or may not possess such knowledge.

We outline a simple hands-on demonstration that teachers can use to illustrate the basic mechanisms underlying random assignment. The demonstration requires no special apparatus, nor does it require previous exposure to topics of probability and statistics. The demonstration is appropriate for students at a variety of educational levels.

The Classroom Demonstration

The demonstration is expedited using several groups of 2 or 3 students, but can also be performed by individuals. The basic demonstration requires that the teacher give each small group a standard deck of playing cards. Each card in the deck functions as an experimental "participant." Like real participants, the playing cards have unique characteristics, or background variables, that might affect a hypothetical dependent variable. These characteristics include the color of the card's print (red or black), whether the card is a face card or number card, and the suit of the card (spade, club, diamond, or heart). If desired, you can devise additional background variables (e.g., the magnitude of the numeric value, odd- vs. even-numbered cards).

The goal of the demonstration is to randomly assign the "participants" to one of two treatment conditions. It would be desirable if the background characteristics are equally distributed between the two treatment groups. There are at least two different ways to simulate this outcome. First, students could draw cards individually from the deck and assign them to treatment conditions using a coin toss. A slightly less time-consuming method of assignment would be to thoroughly shuffle the deck and alternate draws between the two groups. The latter method is advantageous because it ensures equal group sizes, but may leave students confused about the randomization process (e.g., some students may be left with the impression that random assignment involves assigning every other participant to a treatment condition). After the students form the two treatment conditions, they record the number of times each background characteristic (e.g., suit, color) occurred within each treatment condition. For example, students can count the number of spades assigned to each of the two treatment groups. The relative frequency of occurrence can subsequently be compared for each of the background characteristics.

Inevitably, some groups will have nearly flawless randomization, resulting in roughly equal representation of the various background characteristics between the two groups of cards, whereas other groups will have less equivalence. This result is not a flaw in the demonstration, but an accurate re-
reflection of how random assignment operates in the context of experimental studies. Such an outcome serves as an opportunity to discuss the fact that equivalence is, in fact, probabilistic, and how it is possible to obtain disparate groups simply due to chance. In any given sample, random assignment does not guarantee that treatment groups will be equated on every confounding variable. It is important to explain to students that this lack of equivalence is not a failure of randomization, but reflects the natural behavior of a probabilistic process. In fact, instructors can use the binomial distribution to illustrate just how likely small discrepancies (e.g., 11 red cards and 9 black cards) are to occur simply due to chance. This demonstration may help students further understand the probabilistic nature of randomization.

Evaluation

We evaluated the effectiveness of this classroom activity using a 15-item multiple-choice quiz. Ten of these items had content related to random assignment, whereas the remaining five items were filler items related to other research design topics (e.g., random selection, counterbalancing). The resulting quiz scores ranged between 0 and 10, based only on the items related to random assignment.

We administered the quiz to two samples of undergraduate psychology students using a pre–post design. The first sample consisted of 33 participants from an introductory undergraduate statistics and research design course, and the second sample consisted of 21 students from an honors undergraduate statistics and research design course.

The mean scores for the sample of 33 statistics and research design students were 4.57 (SD = 2.13) and 5.88 (SD = 1.89) at the pretest and posttest, respectively. A paired samples t test indicated a significant increase in scores between administrations and a medium effect size, $t(32) = 3.95, p < .001, d = .65$. The mean scores for the sample of honors students were 5.04 (SD = 1.88) and 6.76 (SD = 1.78) at the pretest and posttest, respectively. Again, a paired samples t test revealed a significant increase in scores and a large effect size, $t(20) = 3.87, p < .001, d = .94$.

Discussion

Despite its apparent simplicity, random assignment is an abstract concept for students. This article describes how to demonstrate random assignment using a simple deck of playing cards. If time permits, the instructor can extend the demonstration in a variety of different ways. For example, one of the influences on the randomization process is sample size. It is reasonable to expect that random assignment will become increasingly effective as sample size increases. The instructor can easily incorporate the effects of sample size on randomization into this demonstration by varying the number of cards in each deck. For example, you can give some students a half deck and give other students one, two, or three decks. You can then compare the relative effectiveness of the randomization process across sample size conditions.

You can also simulate quasi-experimental designs with intact groups using this demonstration. When you first hand the decks to the students, they can form two treatment groups by splitting the deck in half. On closer inspection, the students can see that the two groups are dramatically different in some respects; one group may have no clubs or spades, for example. This outcome is exactly the situation that can occur when using intact groups in pre- or quasi-experimental studies. Systematic self-selection may take place based on important background variables (e.g., socioeconomic status, achievement), leading to nonequivalent groups. This extension works particularly well with a new deck of playing cards because new cards often come sorted by suit.

Although more complex, you can also extend the demonstration to compare quantitative differences between the groups. For each card, the numeric value on the face could be treated as a “score” on some psychological variable (e.g., IQ). After students randomly assign the cards to the two groups, they can record both the qualitative and quantitative information about each card within the two groups. For example, they can calculate and compare means between the two treatment groups (e.g., compare the mean of the spades in the first treatment group to the mean of the spades in the second group). You can also use the numeric value of the card to introduce the concept of matching participants based on prior knowledge of their characteristics (e.g., cards with the same numeric value could be randomly assigned to the experimental conditions).

We believe that the exercise and extensions we outlined in this article may be useful for students at different education levels, and we have used the demonstration in undergraduate- and graduate-level research design and statistics courses. Despite the limitations inherent in such a design, we compared pretest and posttest quiz scores in two undergraduate samples and found significant differences. We hope that teachers can use this demonstration in their introductory research design and statistics courses to make an abstract concept more concrete.

References


**Note**

Send correspondence including requests for the quiz to Craig K. Enders, Department of Psychology, Arizona State University, P.O. Box 871104, Tempe, AZ 85287–1104; e-mail: cenders@asu.edu.
Teaching Psychology Students to be Savvy Consumers and Producers of Research Questions

Robert J. Sternberg
Yale University

Students often learn to evaluate research for methodology or results rather than for the problem posed or the question asked. In this article, I suggest the importance of teaching students to be good consumers of research problems and questions, summarize research in this area, and describe activities to cultivate good consumership. I further suggest that students need to learn to become active producers of quality research and research questions because in their lives they will need not only to provide good answers to questions but to ask the right questions.

Many teachers of psychology believe that their students are not as savvy in evaluating research as their training should enable them to be (Sternberg, 1997b). Why? Psychology courses, especially at the introductory levels, often tend to emphasize learning or even rote memorization of facts. The kinds of assessments that instructors use may encourage students to remember the research they read rather than process it more deeply.

Of course, there are many excellent critical-thinking texts available to students (e.g., Bell, 1995; Halonen, 1995; Levy, 1997; Mayer & Goodchild, 1995; McBurney, 1996; Smith, 1995; Stanovich, 1996; Tishman, 1997). Some professors assign sources such as these whereas others do not. The emphasis in these texts tends to be more on thinking about problem solving than about problem formulation, that is, about coming up with worthwhile ideas in the first place (Arlin, 1975/1976). In this article, therefore, I help supplementing rather than replace the perspective such books bring to bear on teaching higher order thinking in psychology. It is one attempt among others to bring a triarchic perspective on human abilities into the psychology classroom. According to the triarchic theory, memory is important, but the abilities needed to analyze, synthesize, and apply the learned information are keys to intelligent thinking (Sternberg, 1985; Sternberg, 1997a). In this article, I suggest that students become synthetic thinkers about research questions and problems as well as critical thinkers about research answers and solutions, always keeping in mind the practicalities of research.

Activities to Promote Good Consumers

I have tried a number of classroom activities that help students to become good consumers of research questions. These activities can be used at any level of teaching psychology.

1. Simulation of journal refereeing. I ask upper-level undergraduate and graduate students to simulate being referees for journals. I show them in advance some of the

References


Notes


2. I thank Randolph Smith and four anonymous reviewers for their helpful suggestions on previous versions of this article.

3. Send correspondence to Jane P. Sheldon, Department of Behavioral Sciences, University of Michigan–Dearborn, Dearborn, MI 48128; e-mail: jsheldon@umich.edu
forms editors and referees use to evaluate manuscripts and go over with them these and other criteria referees use in evaluating articles. I can then evaluate the students’ work in terms of their ability to use the criteria successfully. I emphasize the importance of being constructively critical and of focusing on how well the research is done and whether the research was worth doing in the first place. Students are able to do these tasks quite well. They often find flaws in designs of studies, interpretations that do not follow directly from the results, studies lacking in theoretical motivation, and so forth. The best way ultimately to produce a better pool of referees, as well as of research consumers in general, is through helping students learn how to referee research.

2. Evaluation of why studies are in textbooks or lectures. I ask upper-level undergraduate students to reflect on research described in textbooks or in lectures and to discuss or write about why it is there. In this activity, students come up with a variety of responses. I guide a class discussion to focus on the more insightful responses such as “Many people followed up on the study with their own studies,” “It changed our way of thinking about psychology,” and “A lot of other people cite it too.” In this way, students come to understand not only what a study found, but why it was done and why an instructor or textbook writer thought it was worthy of citation.

3. Evaluation of primary versus secondary descriptions of studies. Students often assume that the way that books and teachers represent research to them is very close to the way the research was done. Sometimes it is not (e.g., Prytula, Oster, & Davis, 1977). Becoming a good consumer of research problems means understanding why and how a study was done and comparing its presentation in the original source to its representation in a secondary source. Even beginning students can learn through this activity that a savvy consumer of research cannot automatically accept the accuracy of an account of research in a secondary source.

These activities develop good consumers of research and research questions, but consumption is only half the story. The other half of the story is production.

Producing Research and Research Questions

It is not enough to appreciate good questions that others ask. Students need to learn to ask good questions themselves—to become proactive and not just reactive thinkers. Even students who learn to be good consumers of research may never develop the skills they need to produce the work they are so ready and willing to criticize. A balance is needed.

Activities to Promote Good Producers

Instructors can use several classroom activities to promote good producers of research.

1. Generating and evaluating alternative studies. As a class exercise, I ask students at all levels to generate and then evaluate some research topics that may be of interest to them and that are relevant to the course. I indicate that if we chose a class project, we would have time to do only one of the studies. I ask them to generate and apply various criteria for deciding which study to conduct. I have found that students generate criteria quite similar to those generated by professional psychologists (Shadish, 1989; Sternberg, 1992, 1993; Sternberg & Gordeeva, 1996; Sternberg, Hojjat, Brigockas, & Grigorenko, 1997).

For example, they raise issues of how interesting and how important the study is, whether the class has the resources to conduct the study, and whether it is likely to work. These are the same issues students would need to address in planning their own research. The activity also gives them a chance to generate their own ideas.

2. Small-group projects from start to finish. In courses of modest enrollment, small-group projects provide a way for students to learn not just the process of doing research, but the process of collaboration with others in this research. In these projects, I not only have students do the research together, but decide what they want to do, how they want to do it, how they will divide the research responsibilities, and how they will divide the presentation responsibilities. Students learn from each other how to generate and do research and also learn their strengths and weaknesses as collaborators so they can improve these collaborative skills.

3. Question generation. Students can generate rather than answer questions. For example, one might form a Jeopardy game where students read a secondary account of a study and try to figure out what question it asks. They then can read the original study to see whether they have inferred the question correctly, based on the secondary-source description. Or one might describe an interaction among people and ask what interesting psychological questions emerge from the interaction. Students can learn to ask good questions by gaining experience in asking them.

Conclusions

Students can and should learn not only about the contents of research, but also about how to generate and evaluate research. To the extent that students now learn about generation and evaluation, they tend to learn about answers and solutions rather than about questions and problems. Instructors need to right this imbalance. They can improve instruction by teaching students directly about criteria for evaluating research in general and research problems in particular. Instructors can also engage students in constructive activities that help them learn firsthand how to formulate as well as evaluate such problems. In this way, students will become the kinds of savvy consumers who, as citizens, will better appreciate the value of research about which they hear. If they become psychologists, they will be-
References

Sternberg, R. J. (1997b). What do students most need to learn about research in psychology? APS Observer, 14, 19.

Notes

1. Preparation of this article was supported under the Javits Act Program Grant R206R50001 administered by the Office of Educational Research and Improvement, U.S. Department of Education. The findings and opinions expressed do not reflect the positions or policies of the Office of Educational Research and Improvement or of the U.S. Department of Education.
2. Send correspondence to Robert J. Sternberg, Department of Psychology, Yale University, Box 208205, New Haven, CT 06520–8205; e-mail: robert.sternberg@yale.edu

Effects of Day Care and Maternal Employment: Views From Introductory Psychology Textbooks

Claire Etaugh
Joseph Cohen
Myra Cummings-Hill
Michelle Massey
Kelly Selchow Detweiler
Bradley University

We analyzed introductory psychology textbooks from 1970 to 1997 for the inclusion and treatment of the effects of day care and maternal employment on children. Coverage increased substantially over the 28 years. Recent texts were less likely than earlier texts to express negative views about maternal employment and day care and were more likely to present positive or evenly balanced views.

Approximately 75% of married women with school age children and 60% of those with preschoolers are employed, and college-educated mothers are especially likely to work (U.S. Bureau of the Census, 1997). Given these data, knowledge of the impact of maternal employment and day care on children’s development has great relevance for college students. A potentially important source of information about these topics is the introductory psychology course, one of the most highly enrolled undergraduate courses. Introductory psychology textbooks may influence deeply students’ beliefs and knowledge of an array of subjects. These texts can play a major role in conveying facts and attitudes about day care and maternal employment, which may influence students long after graduation. Our study examined the treatment of day care and maternal employment in introductory psychology textbooks published between 1970 and 1997.

Method

Sample

Two pairs of senior psychology majors searched the indexes of 115 introductory psychology textbooks for the following key words: child care; day care; latch key children; maternal employment; mother, employed; mother, working; employed mother; working mother; parental influence. They also searched the developmental psychology chapter. Each pair evaluated and rated approximately half the texts.

Ratings of Five Dependent Variables

Raters tabulated the number of pages in each text devoted to day care and maternal employment, considering 50 standard-length (11.5 cm) lines equivalent to one page and counting fractional pages as one page. They converted texts...
Using Daily Horoscopes to Demonstrate Expectancy Confirmation

Geoffrey D. Munro
St. Mary’s College of Maryland

James E. Munro
Edinboro University of Pennsylvania

Prior expectations can lead a person to process information in a biased manner such that the expectations are confirmed. Despite its prevalence in everyday judgments, people lack insight into expectancy confirmation processes. The classroom demonstration we describe uses daily horoscopes to demonstrate the powerful effect that expectations can have on judgments. More students reported that the events of a recent day most closely matched the prediction for their particular astrological sign when the astrological signs were presented along with the predictions than when they were not. Students reported that the demonstration was interesting, and 95% of the students correctly answered a question regarding the phenomenon on a subsequent examination.

Humans have difficulty perceiving the world in a truly objective manner because their expectations influence their judgments. A wealth of research has shown that information processing biases can lead to confirmation of expectancies (Darley & Fazio, 1980; Merton, 1948; Rosenthal & Jacobson, 1968). Expectancies influence the perceptual encoding of information (von Hippel, Sekaquaptewa, & Vargas, 1995), the retrieval of information (Snyder & Uranowitz, 1978), and one’s behavior (Snyder, Tanke, & Berscheid, 1977). Finally, when one has a hypothesis, that person often searches for confirming rather than disconfirming information leading her or him to find evidence that supports the hypothesis (Snyder & Swann, 1978).

Psychology instructors, of course, understand how expectancy confirmation can influence judgments. They often present to students intriguing research examples detailing the biased processing of participants in some distant and past experiment. However, although it is easy to understand how expectations affect participants in these experiments, it is sometimes difficult to appreciate the ease with which everyday judgments can be affected. Indeed, the power of expectancy confirmation comes from this lack of insight. Thus, the instructor’s task is to provide students with insight into how expectancies can influence everyday judgments.

A topic that has been used with considerable success in demonstrating a variety of scientific methods and psychological concepts is the scientific inquiry of astrology (Franknoi, 1988; Jones & Zusne, 1981; Kelly, 1989; Ward & Grasha, 1986). Although astrology has not been used to demonstrate expectancy confirmation, it might be particularly appropriate for that task. When daily horoscopes are compared to actual daily life events, it seems likely that a person’s expectations that horoscopes are accurate might lead that person to find confirming evidence. The demonstration discussed in this research has two advantages that are directly related to the problems associated with expectancy confirmation. First, many students will have the opportunity to experience expectancy confirmation firsthand, and all students will directly observe its effects. Second, as a consequence of observing expectancy confirmation, students should gain an appreciation for the biasing effects of expectancy confirmation in scientific research and everyday judgments.

Method

Preparation

Two days before the class, the instructor collected daily horoscopes from a newspaper and prepared three overhead transparencies. The first, based on the horoscopes for the day that was 2 days before the class, included the zodiacal sign, the range of birth dates that correspond to that sign (for those unfamiliar with their zodiacal sign), and the predictions for that day’s events. The second was identical to the first except that it was based on the horoscopes for the day that was 1 day before the class. The third, also based on the horoscopes for the day that was 1 day before the class, included only the predictions for that day’s events in scrambled order labeled with the numbers 1 through 12 rather than zodiacal signs and birth dates. If daily horoscopes are not available, personality descriptors for each of the zodiacal signs could be substituted (although some students may know the traits associated with each sign).

Procedure

To introduce the demonstration, the instructor feigned enthusiasm for the use of horoscopes as a means of predicting life events. The instructor then posed the question as a scientific hypothesis that should be tested empirically. He explained to the students that one way to test for accuracy is to compare the horoscopes’ daily predictions to the actual
events experienced on that day. Students thought about what happened to them 2 days before. Then, the instructor presented the first overhead. The task of the students was to find the prediction (out of all 12 horoscopes) that best described what happened to them 2 days before. Once they found a prediction, each student passed forward a small slip of paper with their actual zodiacal sign and the sign of the best prediction of what happened to them 2 days before. The instructor then read each slip to the class to determine how many students thought that the best prediction also happened to be their actual zodiacal sign. This number was coded as the number of matches. Depending on the size of the class and the amount of time available, this step of the demonstration could also be completed by a show of hands. A percentage was computed by dividing the number of matches by the total number of students in the class and was then compared to the percentage one would expect to find by chance responding (8.3%).

After simulating excitement about the potential for astrology to predict life events at better than chance levels, the instructor suggested that one more test was necessary. He then presented the third transparency, which showed only the order-scrambled daily event predictions without the accompanying birth dates or zodiacal signs. This time, students found the one prediction that best matched the events of 1 day before the class. Once the students identified a prediction, each student passed forward a small slip of paper with their actual zodiacal sign and the number of the prediction that best fit what happened to them 1 day before. The instructor then presented the second transparency, which included the actual predictions from the previous day as well as the accompanying zodiacal sign. The instructor read each slip to the class to determine (by referring to the second transparency) how many students thought that the best prediction also happened to be their actual zodiacal sign. Again, students could be asked to perform this step on their own with the number being determined by a show of hands. After computing a percentage, it was compared to the percentage one would expect by chance responding.

Results

Demonstration Results

The instructor collected the data from four recent introductory psychology sections. When students indicated that the prediction that best fit the events of the day was also the prediction for their actual zodiacal sign, it was recorded as a match. When students indicated that the prediction that best fit the events of the day was not the prediction for their actual zodiacal sign, it was recorded as a mismatch. The numbers of matches and mismatches were then entered as the observed data in a chi-square where the expected chance data points were 1 match for every 11 mismatches.

When the predictions were labeled with zodiacal signs and birth dates, students reported matches (13%) at a level not significantly different than the 8.3% expected by chance responding, $\chi^2(1, N = 128) = 3.58, p > .05$.

Student Evaluations

To evaluate the demonstration, students from two of the introductory psychology sections responded to a short questionnaire following the demonstration and a discussion of how the demonstration was relevant to the topic of the day. The questionnaire included four rating scale items and two open-ended items. Students responded on a 5-point scale with 1 (strongly agree) and 5 (strongly disagree) as the endpoints and 3 (neutral) as the midpoint, to the following items:

- This exercise was an interesting learning experience.
- Doing this exercise was a valuable way to learn.
- This exercise helped me to understand how expectations can influence judgments.
- Would you recommend this exercise for future classes?

The two open-ended items were as follows:

- What specifically did you learn through the exercise?
- Please write any additional comments below.

The instructor asked the students not to write their names on the questionnaires and assured them that their responses would be anonymous.

For the four rating scale items, the means computed from the 73 students were all between 1 (strongly agree) and 2 (agree: interesting $M = 1.82$, $SD = .56$; valuable $M = 1.67$, $SD = .58$; understand $M = 1.61$, $SD = .57$; recommend $M = 1.90$, $SD = .63$). In fact, no student answered disagree or strongly disagree to any of the questions. In responding to the open-ended question assessing what students learned as a result of the demonstration and discussion, 67 of 73 students (92%) clearly expressed the idea that expectations can influence judgments. Additionally, on an exam given several weeks later, students listed the weaknesses of observational research (the demonstration was originally presented to the students in this context). In responding to the question, 69 of 73 students (95%) correctly answered that one weakness was that the expectations of the observer could influence what observations were reported. Of course, the demonstration is only one of many sources of learning (including textbook readings and reviews of content material) that may have contributed to the students’ examination performance.

Discussion

The demonstration appears to be an effective technique for teaching students about expectancy confirmation. Students found it to be an interesting and valuable way to learn. Moreover, on two separate assessments after the activity, a vast majority of students demonstrated their knowledge of the expectancy confirmation phenomenon. Furthermore,
the demonstration can be used in many different psychology courses to illustrate a wide variety of theoretical and applied examples of expectancy confirmation including the weaknesses of observational research, the double-blind design, the self-fulfilling prophecy, and the fallacy of personal validation of personality assessment (Forer, 1949; Hyman, 1989). Although the demonstration gives students firsthand experience and observation into expectancy confirmation, it is also important for students to interpret the findings after the demonstration is conducted. Specifically, students need to make the connection between the demonstration results and the tendency for people to search for and find expectancy-confirming information. After a firm understanding of expectancy confirmation becomes apparent, a number of potential routes for discussion could follow. First, the instructor could generate a discussion focusing on the different processes by which expectancies are confirmed (e.g., encoding bias, self-fulfilling prophecy, etc.). In a more advanced class (e.g., social psychology), the class could debate whether these processes are driven by wants and desires or by pure information processing biases. Second, the instructor could ask students to generate other real-world situations in which expectancy-confirmation processes might apply. Third, in a research methods course, the instructor could engage students in a discussion regarding the advantages and disadvantages of within-subjects and between-subject designs given that the demonstration makes use of the former whereas an actual experiment would most likely make use of the latter. Finally, one could use the demonstration as a starting point for an investigation into pseudoscientific claims (e.g., the ability of palm readers and psychics to accurately assess personality) and into speculations as to how those claims can best be investigated scientifically (Hyman, 1989).

References


Notes

1. We thank Lucy Bohne, Jennifer Grieme, Terell Lasane, Richard Platt, and several anonymous reviewers for their helpful comments on this article.

2. Send correspondence to Geoffrey D. Munro, Department of Psychology, St. Mary’s College of Maryland, 18952 East Fisher Road, St. Mary’s City, MD 20686; e-mail: gdmunro@osprey.smc.edu.
The Mind as Black Box: A Simulation of Theory Building in Psychology
Carolyn Hildebrandt and Jennifer Oliver
Teaching of Psychology 2000 27: 195
DOI: 10.1207/S15328023TOP2703_06

The online version of this article can be found at:
http://top.sagepub.com/content/27/3/195
METHODS AND TECHNIQUES

The Mind As Black Box: A Simulation of Theory Building in Psychology

Carolyn Hildebrandt
University of Northern Iowa

Jennifer Oliver
Rockhurst University

This activity gives students a collaborative, hands-on experience in theory building. Using the metaphor “the mind is a black box,” students work together in small groups to discover what is inside a sealed, black, plastic box. In all, 63 undergraduate and 11 graduate students evaluated the activity. Students reported that they enjoyed the activity and that it helped them learn more about the development of scientific theories; the existence of conflicting theories; and the value of logical thinking, imagination, and social collaboration in the process of scientific investigation.

Psychology students often do not understand why their professors spend so much time describing, comparing, and evaluating different theoretical approaches to the study of the discipline. “Why do you keep talking about theories?” they ask. “Why don’t you just tell us what’s right?”

Students who ask such questions are unaware of the complexity of problems pursued by psychological researchers and the challenges involved in resolving theoretical debates (King & Kitchner, 1994). Many questions in the physical, biological, and social sciences are difficult to answer because the phenomena in question are not directly observable. These problems are “black box” problems because the structures and processes are hidden (as if in a black box) and must be inferred from other sources (Nairne, 1997). Some of the more important black box problems of contemporary psychology are memory, logico-mathematical thinking, and language development.

Introductory and advanced psychology texts often have a section on methods of scientific inquiry used to build theories. This section usually includes definitions of theory, hypothesis, the scientific method, and a brief overview of the main theoretical perspectives relevant to the field. Most instructors ask students to read this section and then provide a lecture or a discussion on the topic. However, this approach often leaves students with little more than a series of facts that are difficult to understand and apply in a meaningful way.

The activity described here helps students learn more about theories and methods of inquiry by providing a collaborative, hands-on experience in theory building. It is one of a number of problem-solving activities that can help students understand the process and experience of scientific reasoning and investigation (e.g., Doolittle, 1995; Hatcher, 1990; Schlenker, 1993; Stadler, 1998). Using the metaphor “the mind is a black box,” students work together in small groups to discover what is inside a sealed, black, plastic box. After comparing their results with those of other groups with the same box, they conduct additional investigations. This activity can help introductory and advanced students understand the need for theories in psychology and gain insight into how psychologists build, test, and refine those theories. Instructors can refer to this activity throughout the course to show how various psychological constructs (e.g., the cognitive structures proposed by Freud, Chomsky, or Piaget) can be conceptualized as answers to the question, “What is in the black box?” They can also use the critical thinking questions introduced in this activity to help students evaluate conflicting theories in psychology.

Conducting the Activity

This activity was most effective when conducted during one 90-min or two 50-min sessions. Prior to class, students in our introductory classes read the section in their textbook on psychological theories and scientific methods. For these students, we adopted the general definition of a theory as “a coherent set of ideas that helps organize data and make predictions” (Santrock, 1997, p. 35). Students in our graduate classes read additional chapters and articles (e.g., Miller, 1993, chap. 1) and used this activity as a prelude to discussing different types of theories (e.g., model, deductive, inductive, and functional theories).

On the day of the activity, the instructor divided the students into small groups (2 to 3 people) and gave each group a plain, black videocassette box. Each box contained something movable (e.g., a marble) and something immovable (e.g., a piece of Styrofoam glued to the inside of the box). The boxes were sealed with tape and numbered according to their contents (e.g., all boxes with a 1 on the outside had the same contents). For small classes (12 to 15 people), we used only one type of box; for large classes (over 70 to 80) we used four or five.
The task for each group was to draw a picture of the inside of the box. Students were free to manipulate the box but not to open it. As they worked together, the instructor asked them to reflect on the following questions: (a) What did you do to discover what was inside your box? (b) What roles did the people in your group play during the investigation? (c) How could you learn more about what is inside your box? As each group finished its investigation, a representative went to the board and drew a picture of what the group thought was inside its box.

After all groups had supplied a drawing, the instructor began the discussion by commenting on the variety of drawings made by the class. (The range of responses is usually quite wide; it is rare to find two drawings of the same box that are exactly alike). Next, representatives of different groups explained their drawings and described how they came to their conclusions. When groups with the same type of box had different theories of what was inside, the instructor encouraged them to ask, “Why? How is it that we can have the same boxes yet so many different drawings of what is inside?” The instructor used the following questions to facilitate critical thinking and further discussion.

1. “Could we be starting with different assumptions?” Some groups assumed that because the boxes look like videocassette containers, they should have plastic spools inside and did not bother to test this assumption. (We removed all the spools in ours.)

2. “Could we be asking different questions?” Some groups tried to identify the moving objects but did not ask if there was anything else inside that did not move.

3. “Could we be asking the same questions but be using different methods?” Some groups looked only for evidence that confirmed their hypothesis, whereas other groups looked for disconfirming evidence as well.

4. “Could we be asking the same questions, doing the same experiments, getting the same results, but be interpreting our data differently?” Two groups agreed that they heard a “clunk, clunk” sound in their box but differed in their interpretation of it. One group claimed that there were two marbles, each of which went “clunk” as it hit the side of the box. The other group contended that there was only one marble that went “clunk, clunk” as it bounced off the side of the box.

When groups working on the same type of box disagreed, the instructor challenged them to think of new ways to test the adequacy of their theories. They could do so by designing new experiments and collecting new data (taking care to collect both confirming and disconfirming evidence for their theory) or by reformulating the theory itself. During this part of the activity, some groups decided to go back to their box and try new experiments on it; others elected to trade boxes with other groups, replicate one another’s experiments, and do new experiments together.

At the end of the session, the instructor reviewed the basic steps of the scientific method using examples from the activity. She noted that the first step of the scientific method is to identify a problem. In this case, the problem was what was inside the black box. The next step is to collect data, which most groups did by tilting and shaking their box while attending to the sounds it made. The next step is to analyze and interpret the data, which most groups did several times during the investigation, using their insights to fuel additional investigations. The final step is to compare one’s findings to those of other researchers in the field. Students did this by describing their research methods and sharing their drawings with one another. In the face of new or contradictory evidence (which was not lacking in this activity), there are several courses of action open to the researcher—to revise the theory, discard the theory, conduct additional investigations, or find another problem to pursue.

After discussing the scientific method, the instructor expanded the discussion to include the following questions: (a) What are the similarities and differences between the black boxes and the human mind? (b) What are the similarities and differences between the methods that the class used with the black boxes and those that psychologists use to study the human mind?

The instructor referred to questions raised during this activity throughout the semester when comparing and contrasting competing psychological theories. The black box metaphor was especially useful in helping students understand debates between competing theories (such as behaviorist and innatist accounts of language development or psychoanalytic, structural developmental, and social learning accounts of social and moral development).

Student Evaluations of the Activity

Two groups of students evaluated the activity. One group consisted of 63 undergraduates enrolled in an introductory developmental psychology class, and the other consisted of 11 graduates enrolled in a master’s level seminar in developmental psychology. The first author taught both classes.

The students completed an anonymous questionnaire consisting of eight questions. Four of the questions included a 5-point rating scale and room for written comments (see Table 1). The remaining questions included space for short, written answers.

Evaluations of the activity were generally positive. Mean responses to the quantitative questions ranged from 4.00 to 4.46 (see Table 1). In response to the qualitative questions, both graduate and undergraduate students reported that the activity was fun, intriguing, mind challenging, and totally suspenseful. Many students commented on how much they enjoyed working with other people and sharing different points of view. In response to the question “What did you learn from the activity?” students wrote that they learned about “different ways to go about theory building” and “the importance of imagination in science.” They also learned that “brainstorming and hypothesizing can actually work” and that “theories contradicting each other is [sic] good because it keeps us wondering, questioning, and always learning.”

One complaint raised by a large number of students was that they never got to open the boxes and see what was inside. They did not like the ambiguity of the situation and wanted to know conclusively whether their drawings were right or wrong. In contrast, a smaller number of students
said that they preferred not being told what was in the box and appreciated the opportunity to ponder its contents and reflect on other theories that have no definite or agreed-on answer. In fact, one student maintained that keeping the boxes closed was an important part of the activity and suggested that the professor should “never reveal the contents” to the class.

When students ask what is inside the boxes, our policy is to say, “I don’t remember” or “I don’t know,” with a knowing twinkle in our eyes. We believe that if we told them what is inside, they might be tempted to think that there is a right answer to every question in psychology. This conclusion would run counter to the purpose of the activity, which is to simulate problems where one cannot simply “open the box and see what’s inside.”

References


Notes

1. This article was presented at the fourth Midwest Institute for Teachers of Psychology, Glen Ellyn, IL, March 1997; and the American Psychological Association Conference, San Francisco, August 1998.

2. The first author thanks Paul Ammon for introducing us to this activity in 1980. (Paul, in turn, thanks Lawrence Lowrey and Robert Karplus for introducing the activity to him.) Thanks are also due to Rheta DeVries, Andrew Gilpin, Roger Sell, David Whitsett, and Jack Yates for their helpful comments on the article and activity.

3. Send correspondence to Carolyn Hildebrandt, Department of Psychology, 346 Baker Hall, University of Northern Iowa, Cedar Falls, IA 50614–0505; e-mail: carolyn.hildebrandt@uni.edu.

Table 1. Student Evaluations of the Black Box Activity

<table>
<thead>
<tr>
<th>Questions</th>
<th>Undergraduatesa</th>
<th>Graduatesb</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Did you find the activity interesting?c</td>
<td>4.00 .83</td>
<td>4.36 .67</td>
</tr>
<tr>
<td>2. How useful do you think this activity was in describing theory-building/scientific method?d</td>
<td>4.20 .78</td>
<td>4.36 .51</td>
</tr>
<tr>
<td>3. How well did the activity help explain the existence of conflicting theories in psychology?e</td>
<td>4.29 .77</td>
<td>4.46 .69</td>
</tr>
<tr>
<td>4. How highly would you recommend this activity to other professors?f</td>
<td>4.24 .86</td>
<td>4.27 .79</td>
</tr>
</tbody>
</table>

*a n = 63. b n = 11. c Based on a scale ranging from 1 (not at all interesting) to 5 (very interesting). d Based on a scale ranging from 1 (not at all useful) to 5 (very useful). e Based on a scale ranging from 1 (not very well) to 5 (very well). f Based on a scale ranging from 1 (would not recommend) to 5 (would highly recommend).
dent reported being home watching television when her mother informed her of Princess Diana’s death. In the follow-up probe, the student reported being at her boyfriend’s house studying for an exam when he came home with the news. In both of these accounts, the student reported location, ongoing activity, and informant information. Afterwards disappeared in the second probe, and the other three categories of information disagreed.

The project led to lively discussions. Students were pleased and eager to describe their consistent memories as well as their memory discrepancies. One of the most useful aspects of the project is that it provides the class with self-generated examples of many of the concepts covered in an introductory course. For instance, when we want to discuss omission or elaboration, we ask students to search their probes for examples of accounts in which they preserved the gist but added or dropped details. When we discuss schemas, we ask students to search for inconsistent accounts and to speculate on how they may have used an existing schema to reorganize their memories. When examining the effects of emotion, we ask students who had experienced high levels of emotion at the time of encoding to compare their memories with students who had reported little emotion. The same type of discussion can take place when introducing rehearsal. We have found this exercise to be particularly successful for generating examples and illustrating the roles of emotion, rehearsal, consolidation, elaboration, omission, reconstruction, and schemas.

We assessed the effectiveness of this project in two ways: (a) by examining the impact on classroom test scores and (b) by examining course evaluations. We compared scores on examination questions concerning episodic and flashbulb memories (M = 67.84, SD = 22.97) to a sample of factual and conceptual questions concerning other memory phenomena on the same examinations (M = 63.71, SD = 18.09). Students scored higher on the episodic and flashbulb memory questions, t(152) = 2.20, p = .03. In contrast, students enrolled in sections that did not use this project tallied similar scores on these two sets of questions (M = 63.44, SD = 18.34 vs. M = 63.38, SD = 21.97), t(241) = 0.03, p = .97. We also compared scores on the episodic and flashbulb memory questions to performance on those same questions in course sections that did not use this project. Students who had completed this project (M = 67.84, SD = 22.97) scored higher on those questions (M = 63.38, SD = 21.97), t(393) = 1.98, p < .05. Thus, although the effect size is small, it does appear that recording their memories helped the students learn about episodic and flashbulb memories.

We solicited evaluations of the project and the course from students at the end of the semester. Students rated the course as a 4.4 and the project as a 4.6 on a 5-point scale ranging from 1 (poor) to 5 (outstanding). Written comments were also uniformly positive. Students described the project as “intriguing” and “interesting.” They were especially pleased to be “able to analyze ‘real data.’” Students also reported that they believed the project helped them to better understand memory processes.

Overall, we believe this project is rewarding and instructive for both faculty and students. The project improved student involvement and learning. It required just a few minutes of class time to collect the memory probes and as much time in discussion as the instructor desires. It is a project that can be performed every semester. Many events occur that might trigger vivid memories. For example, recent flashbulb memory research (for a review, see Conway, 1995) has examined memories for the assassination attempt on President Reagan, the San Francisco earthquake, and the Los Angeles earthquake, among others. Although events such as these do not occur at the onset of every semester, we have been able to use the same event for several semesters until another unexpected and emotional event occurs. Finally, we believe that this project is flexible enough to be useful in a number of courses. In introductory psychology we have used it as a vehicle for discussion of memory phenomena. We have also used this project in upper level learning and memory courses. Here, we require more extensive data analysis, a literature review, and completion of a laboratory paper. As a result we have been able to impart concepts in research design and analysis as well as teach about memory phenomena.

References


Note

Send correspondence to Mark Sudlow Hoyert, Psychology Department, Indiana University Northwest, Gary, IN 46408; e-mail: mhoyert@iunhw1.iun.indiana.edu.

Demonstrating the Concept of Illusory Correlation

Jay W. Jackson
Indiana University-Purdue University

The concept of illusory correlation helps explain how some social stereotypes are partly due to basic cognitive processes. I describe a classroom activity that essentially replicates, in a simplified format, a classic study of the illusory correlation effect. In addition to effectively demonstrating the phenomenon of illusory correlation, this activity provides a stimulus for discussing related social psychological issues and methodological procedures.

Research has demonstrated that people pay special attention to distinctive or rare events and recall such events with...
relative ease (Hamilton & Gifford, 1976; Hamilton & Sherman, 1996). For example, a cashier at a fast food restaurant is more likely to remember a person who paid with a $100 bill than a person who paid with a $5 bill. When two distinctive or rare events co-occur, people tend to notice that relation more readily, encode the information more effectively, and recall such relations with greater ease. Consider the perceived relation between rain and weekends. Both of these events are rare and distinctive in that most days are rain free and most days are weekdays. However, when these two relatively rare events coincide, people are more likely to attend to the relation and subsequently recall such “rainy weekends” with relative ease (Matthews, 1996). On the other hand, when the weather or the day of the week is more typical, the relation between these events does not lead to special attention or encoding. Consequently, people often subjectively overestimate how often two distinctive events occur together. This tendency has been dubbed the illusory correlation because it involves perceiving a relation that does not exist or is weaker in reality than perceived (Garcia-Marques & Hamilton, 1996). This cognitive process can influence the content of stereotypes. For example, it partly explains why many White Americans overestimate the rate at which African Americans engage in criminal activity (both are distinct or rare events).

I have conducted a classroom demonstration of the illusory correlation effect that “replicates” a classic experiment conducted by Hamilton and Gifford (1976, Study 1). Students read about people from a numerical majority group (Group A) and a numerical minority group (Group B). Both groups are associated with more desirable behaviors than undesirable behaviors, and the proportion of desirable and undesirable behaviors is the same for both groups (a 9:4 ratio). Because both Group B and undesirable behaviors are rare events, students tend to form illusory correlations and judge Group B less favorably than Group A. Just as researchers have found the effect to be highly reliable across various laboratory settings (Hamilton & Sherman, 1996; Lieberman, 1999), I have found the results to be quite reliable in the classroom. This demonstration helps illustrate the role that basic cognitive processes play in the development of social stereotypes.

The demonstration takes about 20 min and requires a total of 39 slides, each containing a description of a Group A member or Group B member engaging in either a desirable or an undesirable behavior (see Table 1). The descriptions may be presented using computerized slides, overheads, traditional slides, index cards, or other means of presenting written visual stimuli. Reading the descriptions also works as well.

Following the procedures used by Hamilton and Gifford (1976, Study 1), I inform the students that they will see a series of statements, each describing a person performing some type of behavior, and explain that each person described will belong to one of two groups, simply referred to as Group A and Group B. I ask students to read (or listen to) each statement carefully as it is presented. I then present each statement, in a random order, for about 8 sec. After they have seen all the statements, the students complete a rating form. On this form, students indicate their “first impressions” by judging each group, on 7-point Likert scales ranging from 1 (strongly disagree) to 7 (strongly agree), as being generally popular, lazy, unhappy, intelligent, honest, irresponsible, helpful, strong, and unpopular. Second, the students provide frequency estimates regarding the behaviors exhibited by the members of each group (from Hamilton & Gifford, 1976, Study 1):

<table>
<thead>
<tr>
<th>Group A desirable</th>
<th>Jon, a member of Group A, visited a sick friend in the hospital.</th>
<th>Bill, a member of Group A, is rarely late for work.</th>
<th>Andrew, a member of Group A, planted seedlings in the park.</th>
<th>Josh, a member of Group A, finished his homework on time.</th>
<th>Ken, a member of Group A, helped a child lost in the supermarket.</th>
<th>Kevin, a member of Group A, donated his old clothes to charity.</th>
<th>Jeff, a member of Group A, volunteered to tutor needy students.</th>
<th>Eric, a member of Group A, drove his elderly neighbor to the grocery store.</th>
<th>Colin, a member of Group A, works out to keep himself in good shape.</th>
<th>Scott, a member of Group A, received a promotion at work.</th>
<th>Eliot, a member of Group A, sings in the church choir.</th>
<th>Don, a member of Group A, took a hurt stray dog to the vet.</th>
<th>Craig, a member of Group A, helped his friend move.</th>
<th>Tom, a member of Group A, shared his lunch with a coworker.</th>
<th>Nathan, a member of Group A, took the neighborhood kids swimming.</th>
<th>David, a member of Group A, read a story to his daughter.</th>
<th>Fred, a member of Group A, gave blood to the Red Cross.</th>
<th>Gary, a member of Group A, earned an “A” on his research paper.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group A undesirable</td>
<td>Tom, a member of Group A, cheated on an exam.</td>
<td>Chad, a member of Group A, always talks about himself and his problems.</td>
<td>Alex, a member of Group A, kicked a dog.</td>
<td>Ted, a member of Group A, ran a red light.</td>
<td>Vincent, a member of Group A, forgot about his job interview.</td>
<td>Robert, a member of Group A, talks with food in his mouth.</td>
<td>William, a member of Group A, rarely washes his car.</td>
<td>Ron, a member of Group A, made prank phone calls to his teacher.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group B desirable</td>
<td>Bob, a member of Group B, helped a child.</td>
<td>Henry, a member of Group B, went out of his way to return a lost wallet to the owner.</td>
<td>Lane, a member of Group B, is well liked by his colleagues.</td>
<td>David, a member of Group B, converses easily with people he does not know well.</td>
<td>Mark, a member of Group B, learned how to fly an airplane.</td>
<td>Pete, a member of Group B, is recognized as an excellent musician.</td>
<td>Roger, a member of Group B, repaired his neighbor’s lawnmower.</td>
<td>Keith, a member of Group B, organized a birthday party for his friend.</td>
<td>John, a member of Group B, is considered to be a very dependable worker.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group B undesirable</td>
<td>Allen, a member of Group B, dented the fender of a parked car and didn’t leave his name.</td>
<td>Bruce, a member of Group B, never returns library books on time.</td>
<td>Richard, a member of Group B, yelled at a little boy who bumped into him.</td>
<td>Norman, a member of Group B, often tailgates others when he is driving.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stimuli Materials for Illusory Correlation Demonstration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group A desirable</td>
</tr>
<tr>
<td>Bill, a member of Group A, is rarely late for work.</td>
</tr>
<tr>
<td>Andrew, a member of Group A, planted seedlings in the park.</td>
</tr>
<tr>
<td>Josh, a member of Group A, finished his homework on time.</td>
</tr>
<tr>
<td>Ken, a member of Group A, helped a child lost in the supermarket.</td>
</tr>
<tr>
<td>Kevin, a member of Group A, donated his old clothes to charity.</td>
</tr>
<tr>
<td>Jeff, a member of Group A, volunteered to tutor needy students.</td>
</tr>
<tr>
<td>Eric, a member of Group A, drove his elderly neighbor to the grocery store.</td>
</tr>
<tr>
<td>Colin, a member of Group A, works out to keep himself in good shape.</td>
</tr>
<tr>
<td>Scott, a member of Group A, received a promotion at work.</td>
</tr>
<tr>
<td>Eliot, a member of Group A, sings in the church choir.</td>
</tr>
<tr>
<td>Don, a member of Group A, took a hurt stray dog to the vet.</td>
</tr>
<tr>
<td>Craig, a member of Group A, helped his friend move.</td>
</tr>
<tr>
<td>Tom, a member of Group A, shared his lunch with a coworker.</td>
</tr>
<tr>
<td>Nathan, a member of Group A, took the neighborhood kids swimming.</td>
</tr>
<tr>
<td>David, a member of Group A, read a story to his daughter.</td>
</tr>
<tr>
<td>Fred, a member of Group A, gave blood to the Red Cross.</td>
</tr>
<tr>
<td>Gary, a member of Group A, earned an “A” on his research paper.</td>
</tr>
<tr>
<td>Group A undesirable</td>
</tr>
<tr>
<td>Chad, a member of Group A, always talks about himself and his problems.</td>
</tr>
<tr>
<td>Alex, a member of Group A, kicked a dog.</td>
</tr>
<tr>
<td>Ted, a member of Group A, ran a red light.</td>
</tr>
<tr>
<td>Vincent, a member of Group A, forgot about his job interview.</td>
</tr>
<tr>
<td>Robert, a member of Group A, talks with food in his mouth.</td>
</tr>
<tr>
<td>William, a member of Group A, rarely washes his car.</td>
</tr>
<tr>
<td>Ron, a member of Group A, made prank phone calls to his teacher.</td>
</tr>
<tr>
<td>Group B desirable</td>
</tr>
<tr>
<td>Henry, a member of Group B, went out of his way to return a lost wallet to the owner.</td>
</tr>
<tr>
<td>Lane, a member of Group B, is well liked by his colleagues.</td>
</tr>
<tr>
<td>David, a member of Group B, converses easily with people he does not know well.</td>
</tr>
<tr>
<td>Mark, a member of Group B, learned how to fly an airplane.</td>
</tr>
<tr>
<td>Pete, a member of Group B, is recognized as an excellent musician.</td>
</tr>
<tr>
<td>Roger, a member of Group B, repaired his neighbor’s lawnmower.</td>
</tr>
<tr>
<td>Keith, a member of Group B, organized a birthday party for his friend.</td>
</tr>
<tr>
<td>John, a member of Group B, is considered to be a very dependable worker.</td>
</tr>
<tr>
<td>Group B undesirable</td>
</tr>
<tr>
<td>Bruce, a member of Group B, never returns library books on time.</td>
</tr>
<tr>
<td>Richard, a member of Group B, yelled at a little boy who bumped into him.</td>
</tr>
<tr>
<td>Norman, a member of Group B, often tailgates others when he is driving.</td>
</tr>
</tbody>
</table>

| n = 18. | n = 8. | n = 9. | n = 4. |

There were 26 statements about Group A members. In your estimate, how many of these 26 statements described
an undesirable behavior? Please write a number between 0 and 26 in the box below to indicate your estimate.

There were 13 statements about Group B members. In your estimate, how many of these 13 statements described an undesirable behavior? Please write a number between 0 and 13 in the box below to indicate your estimate.

After students complete the rating forms, I collect them and ask for volunteers to help calculate means for each trait rating and for the frequency estimates. In the meantime, the other students discuss what they expect the results to be and why they expect such results. The student volunteers record the results on a transparency that shows the means obtained on each trait for Group A versus Group B members and the percentage of undesirable behaviors estimated for each group. I have consistently found that, as expected, Group B (the minority group) is, on average, rated less favorably on most of the traits than Group A, and students attribute proportionally more undesirable behaviors to Group B than to Group A. The results of two recent demonstrations appear in Table 2. Although I stress the importance and logic of statistical testing, I do not utilize class time to actually perform such tests. Because the effect has been found to be reliable in published research articles (e.g., Hamilton & Gifford, 1976), I prefer to focus on the interpretation of the results and the implications for understanding prejudice and stereotyping.

In discussing the demonstration, I first explain that the demonstration was based on actual research conducted by Hamilton and Gifford (1976, Study 1) and provide an overview of that experiment. I emphasize that Group B members, in both the original experiment and in the classroom demonstration, actually committed undesirable acts at the same rate as Group A members. The basic explanation is that the joint occurrence of two distinctive events attracts attention and can lead to faulty impressions. Illusory correlations may help explain the development of false stereotypes. For example, surveys show that White Americans tend to overestimate the arrest rate of African Americans (Hamilton & Sherman, 1996). The phenomenon of illusory correlation may help, in part, explain this tendency because African Americans constitute a minority and undesirable behaviors are uncommon events.

After the demonstration, I ask the class what other stereotypes might be explained by this concept (characteristics of Asian Americans, religious leaders, the mentally ill, and adolescents are common responses) and how the mass media contribute to these phenomena. For example, when a mentally ill person commits an act of violence (e.g., the shooting of John Lennon or Ronald Reagan), the person's mental condition commands attention, as does the violent act itself. Both are relatively infrequently experienced by the general public, and thus their co-occurrence is especially memorable.

In discussing the possibility that the results were due to negative feelings or beliefs about minorities in general, I point out that Hamilton and Gifford (1976, Study 2) conducted a second study in which the rare behaviors were desirable, rather than undesirable. This change resulted in the minority group (Group B) being rated more favorably on the traits than the majority group.

Of course, the predicted results will not always be obtained for each item and not every student will express the expected bias. Such contrary results provide an opportunity to discuss the probabilistic nature of psychological research and the importance of individual differences. I usually ask the students to discuss why some people might be more prone to the effect than others.

To highlight a methodological issue, I also ask the class why Hamilton and Gifford (1976) used “Group A” and “Group B” in the experiments instead of actual groups. Of course, the reason is to control for previously established associations between groups and traits. Unlike those who judged Groups A and B, people often have preexisting biases for or against real groups. Further research (e.g., Hamilton, Dugan, & Trolier, 1985; Hamilton & Rose, 1980) has shown that preexisting stereotypes can lead people to see correlations that are not there (e.g., doctors who are wealthy, accountants who are timid, librarians who are quiet). Information that is consistent with stereotypes is easier to encode and is more likely to be recalled later (Rothbart, Evans, & Fulero, 1979).

Students generally find this demonstration interesting and valuable in clarifying the concept of illusory correlation. Student evaluations of two recent demonstrations appear in Table 3. In addition to clarifying a cognitive basis of stereotyping,

### Table 2. Mean Trait Ratings and Frequency Estimates for Two Demonstrations of the Illusory Correlation Effect

<table>
<thead>
<tr>
<th>Trait</th>
<th>Session 1a</th>
<th>Session 2b</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Group A</td>
<td>Group B</td>
</tr>
<tr>
<td>Popular</td>
<td>4.64</td>
<td>3.86</td>
</tr>
<tr>
<td>Lazy</td>
<td>2.28</td>
<td>3.41</td>
</tr>
<tr>
<td>Unhappy</td>
<td>3.42</td>
<td>4.94</td>
</tr>
<tr>
<td>Intelligent</td>
<td>5.57</td>
<td>4.14</td>
</tr>
<tr>
<td>Honest</td>
<td>5.35</td>
<td>3.00</td>
</tr>
<tr>
<td>Irresponsible</td>
<td>2.65</td>
<td>5.28</td>
</tr>
<tr>
<td>Helpful</td>
<td>5.17</td>
<td>3.41</td>
</tr>
<tr>
<td>Unpopular</td>
<td>3.07</td>
<td>4.21</td>
</tr>
<tr>
<td>Statements</td>
<td>31</td>
<td>51</td>
</tr>
</tbody>
</table>

Note. Trait ratings were based on a scale ranging from 1 (strongly disagree) to 7 (strongly agree) to group.

For each item and not every student will express the expected bias. Such contrary results provide an opportunity to discuss the probabilistic nature of psychological research and the importance of individual differences. I usually ask the students to discuss why some people might be more prone to the effect than others.

To highlight a methodological issue, I also ask the class why Hamilton and Gifford (1976) used “Group A” and “Group B” in the experiments instead of actual groups. Of course, the reason is to control for previously established associations between groups and traits. Unlike those who judged Groups A and B, people often have preexisting biases for or against real groups. Further research (e.g., Hamilton, Dugan, & Trolier, 1985; Hamilton & Rose, 1980) has shown that preexisting stereotypes can lead people to see correlations that are not there (e.g., doctors who are wealthy, accountants who are timid, librarians who are quiet). Information that is consistent with stereotypes is easier to encode and is more likely to be recalled later (Rothbart, Evans, & Fulero, 1979).

Students generally find this demonstration interesting and valuable in clarifying the concept of illusory correlation. Student evaluations of two recent demonstrations appear in Table 3. In addition to clarifying a cognitive basis of stereotyping,

### Table 3. Means and Standard Deviations on Evaluative Items Across Two Sessions Demonstrating the Illusory Correlation Effect

<table>
<thead>
<tr>
<th>Item</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>I enjoyed this demonstration.</td>
<td>5.78</td>
<td>0.69</td>
</tr>
<tr>
<td>This demonstration helped clarify the concept of illusory correlation.</td>
<td>6.00</td>
<td>0.39</td>
</tr>
<tr>
<td>I would recommend that this demonstration be used in future classes.</td>
<td>5.92</td>
<td>0.47</td>
</tr>
<tr>
<td>This demonstration was interesting.</td>
<td>5.92</td>
<td>0.61</td>
</tr>
<tr>
<td>I will probably tell my friends or family members about this demonstration.</td>
<td>5.42</td>
<td>1.01</td>
</tr>
<tr>
<td>After participating this demonstration, I want to learn more about the nature of stereotypes.</td>
<td>5.36</td>
<td>1.15</td>
</tr>
</tbody>
</table>

Note. N = 37. Responses were based on a scale ranging from 1 (strongly disagree) to 7 (strongly agree).
and prejudice, this demonstration stimulates students to think about other factors that contribute to how people judge individuals and groups.

References


Notes

1. Many thanks to David Hamilton for kindly sharing his research materials and to the anonymous reviewers for their helpful comments on an earlier draft of this work.

2. Send correspondence to Jay W. Jackson, Department of Psychology, Indiana–Purdue University, 2101 Coliseum Boulevard, Fort Wayne, IN 46805–6403; e-mail: jacksonj@ipfw.edu.

“Parenting” Students: Applying Developmental Psychology to the College Classroom

Mary Barnas
Marietta College

In this article, I use Baumrind’s (1971) research on parenting styles to help understand the structure and dynamics of the college classroom. I argue that the way professors view their students affects course syllabi and teaching style. The discussion also reflects the development of my own teaching style and philosophy.

After 6 years, my teaching philosophy and attitude toward my students completely changed. I realized that my years of training as a developmental psychologist were applicable to my teaching. Underlying the changes in my classroom policies were more fundamental changes in my assumptions about the developmental level of my students. This discussion reflects my development as a professor using Baumrind’s (1971) parenting categories as a model.

The Permissive Teacher: “Come to Class if You Want To”

I began teaching at a private, liberal arts institution. I was committed to the notion that my students were adults. I had no attendance policy and no “policing” strategies in place (e.g., pop quizzes) to check whether students kept up with their reading. External contingencies would do little to help them after graduation. Quickly, I became frustrated. Attendance was poor and only a few students participated in the class. After 3 years, I accepted a new position at a somewhat smaller institution where teaching is the primary mission. However, nothing had changed. Attendance and participation were low, and my enthusiasm for this noble profession was reaching a similar level. Still, I held fast to my values. It was only after the students began to ask for the very policies I assumed they resented that I started contemplating fundamental changes. I was shocked to read comments on my evaluations suggesting that I should “have an attendance policy so we will come to class” and “have quizzes so that we won’t put off all of the reading.” One day, I uncharacteristically called on a student in my class and kept questioning him until he at least made an attempt to respond. After class, the student approached me and said, “Dr. Barnas, thanks for doing that. That’s the only way I’m going to get into this class and start really thinking.” The final indication that something may have been wrong with my teaching philosophy came when a student I knew fairly well confided in me that some students were under the impression that I did not care about their investment in my courses because I was lenient about deadlines for assignments.

I began to examine my teaching philosophy and to explore the relation between the developmental literature on parenting styles and my behavior thus far in the classroom. I had been exhibiting a permissive style of teaching. I had few rules and rarely enforced those that existed, not because I did not care, but because I believed the students would be best served by making these decisions for themselves. Baumrind and Black (1967) found that children parented under the permissive style are often unsure of themselves and lack self-control. It seems as though allowing self-imposed limits on immature human beings has the opposite of the intended effect. They flounder and resent authority figures for not providing guidelines. Perhaps although physically mature, these students were not ready to face the adult world of self-imposed deadlines. What next?

Authoritarian Teaching: “Come to Class or Else”

I changed my course policies starting with attendance and then adding pop quizzes. I would not accept late papers. I gave no explanation for these policies but only made it clear I would enforce them. The result was that students were attending my classes in record numbers, and they were more of-
Several authors have increasingly recognized the problem of pseudoscience as a major threat confronting psychology and allied disciplines. We discuss the importance of courses in science and pseudoscience to undergraduate education in psychology and provide (a) a model syllabus for courses in the science and pseudoscience of psychology, (b) a list and description of suggested primary and supplemental texts for such courses, (c) a list of useful educational videos on science and pseudoscience, and (d) suggested Web sites that offer critical evaluations of pseudoscientific claims. Finally, we briefly review the literature concerning the efficacy of courses in the science and pseudoscience of psychology and offer suggestions for future research in this area.

The past decade has witnessed an increasing appreciation of, and concern for, the problem of pseudoscience in contemporary psychology (e.g., Kalal, 1999; Lilienfeld, 1998; McFall, 1991; Tavris, 1998). In recent years, a large number of pseudoscientific or otherwise questionable psychological practices and areas of study have either appeared on the scene or flourished with unabated vigor. Extrasensory perception (ESP), astrology, biorhythms, subliminal self-help tapes, polygraph (“lie detector”) testing, projective measures of personality (e.g., the Rorschach Inkblot Test), New Age psychotherapies (e.g., primal scream therapy, Rolfing), unvalidated herbal remedies for enhancing memory and mood, and suggestive therapeutic techniques (e.g., hypnosis, guided imagery) for recovering purported memories of child abuse and alien abductions comprise only a small subset of such domains in contemporary psychology (for discussions of these and other questionable psychological claims, see Hines, 1988; Shermer, 1997; M. T. Singer & Lalich, 1996). Although pseudoscientific thinking pervades many disciplines, the intrinsic difficulties involved in conclusively falsifying claims concerning human behavior (Meehl, 1978) may render many areas of psychology, particularly the so-called “softer” domains of personality, clinical, counseling, and educational psychology, especially vulnerable to pseudoscience. In addition, pseudoscientific thinking may contribute to the uncritical acceptance of unsubstantiated assertions regarding many purported phenomena outside of the domain of traditional psychology, including unidentified flying objects (UFOs), the Bermuda Triangle, the Loch Ness Monster, and Pyramid Power (Hines, 1988; Randi, 1982). Recent polls of the American public reveal that approximately one half endorse beliefs in ESP, astrology, faith healing, and communication with the dead, and that about one quarter to one third endorse beliefs in ghosts and the lost continent of Atlantis (see Shermer, 1997, 1999). There are approximately 20 times as many astrologers as astronomers in the United States, and more Americans believe in ESP than in Darwinian evolution (Gilovich, 1991). Moreover, with the arrival of the millennium, pseudoscientific beliefs concerning a variety of extraordinary claims, including alien visitation, doomsday scenarios, and ancient prophecies (e.g., the prophecies of Nostradamus), have recently come to the fore (Schafer & Cohen, 1998), and there is suggestive evidence that the prevalence of many pseudoscientific beliefs has increased over the past two decades (Shermer, 1998).

Although pseudoscience and science probably differ in kind rather than degree, most pseudoscientific disciplines can be conceptualized as sharing a covarying set of characteristics (Leahey & Leahey, 1983; Lilienfeld, 1998; for a discussion of the features of pathological science, see also Langmuir, 1989). Among these characteristics are (a) unfalsifiability (Popper, 1959), (b) absence of self-correction (Herbert et al., in press), (c) overuse of ad hoc immunizing tactics designed to protect theories from refutation (Lakatos, 1978), (d) absence of “connectivity” (Stanovich, 1998, p. 116) with other domains of knowledge (i.e., failure to build on extant scientific constructs; Bunge, 1967), (e) the placing of the burden of proof on critics rather than on the proponents of claims (Shermer, 1997), (f) the use of obscurantist language (i.e., language that seems to have as its primary function to confuse rather than clarify; Hockenbury & Hockenbury, 1999; van Rillaer, 1991), and (g) overreliance on anecdotes and testimonials at the expense of systematic evidence (Herbert et al., 2000).

We do not wish to imply that all of the claims we discuss in this article will ultimately be shown to be entirely devoid of.
Psychology teachers should be concerned about pseudoscientific beliefs for several reasons. First, psychology students are almost incessantly bombarded with pseudoscientific claims through their exposure to the entertainment media (e.g., talk shows), the Internet, supermarket tabloids, and "pop" psychology books (Heaton & Wilson, 1995; Lilienfeld, 1998; G. M. Rosen, 1993). Moreover, the increasing accessibility of these information sources makes it likely that undergraduates' pseudoscientific beliefs will continue to flourish in the coming decades. Yet many undergraduates emerge from psychology courses with relatively little ability to critically evaluate pseudoscientific claims (McBurney, 1996).

Second, pseudoscientific psychological beliefs are harmful in several ways (Gilovich, 1991). For example, such beliefs can render individuals vulnerable to the dangers posed by inadequately validated treatments. Psychotherapists who unwittingly implant memories of satanic ritual abuse and alien abduction, for instance, may be creating analogues or even full-blown variants of posttraumatic stress disorder in their clients (Chu, 1998). Therapists who use facilitated communication (Mulick, Jacobsen, & Kobe, 1993) in an effort to elicit language from autistic children are instilling false hopes in parents.

Third, pseudoscientific thinking is a slippery slope (Sagan, 1995a; for a somewhat different view, see B. Singer, 1977). An inability to critically evaluate claims regarding astrology and ESP, for example, may ultimately render individuals incapable of critically evaluating scientific assertions concerning more dire threats to the welfare of the planet, such as environmental hazards (e.g., global warming) and overpopulation (Nickel & Shelton, 1996; Sagan, 1995c).

This heightened concern with the problem of pseudoscience has been paralleled by an increased interest in the teaching of critical thinking skills in undergraduate psychology courses. Teaching of Psychology, for example, devoted a special issue to the topic of critical thinking (Nummedal & Halpern, 1995), and has since featured a number of other articles concerned with the teaching of critical thinking skills (e.g., Griggs, Jackson, Marek, & Christopher, 1998; Marek, Jackson, Griggs, & Christopher, 1998). Although we believe that imparting such skills is extremely valuable, we contend that a fundamental understanding of the differences between science and pseudoscience represents a crucial, yet often neglected, component of critical thinking. As Lawson (1999) noted,

"Psychology students should be able to think critically or evaluate claims, in a way that explicitly incorporates basic principles of psychological science ... psychology majors should be able to judge claims based on a lack of empirical evidence, testimonial or anecdotal evidence, unfalsifiable theories, biased samples, or simple correlational data as weak claims. (p. 207)"

Moreover, there is evidence that critical thinking skills are best learned within the context of a specific discipline rather than in the abstract (McBurney, 1996; Resnick, 1987). As a consequence, teaching students critical thinking skills within the context of specific pseudoscientific claims (e.g., ESP, subliminal persuasion, lie detection) affords them a valuable opportunity to acquire and hone such skills while gaining exposure to topics that they find intrinsically interesting and engaging.

A final advantage of educating students about the differences between science and pseudoscience stems from the theorizing of G. A. Kelly (1955), who observed that many phenomena can be fully grasped only by understanding their opposites. From a Kellyian perspective, students' comprehension and appreciation of science can be enhanced by a thoughtful examination of pseudoscience and its characteristics (see also Wesp & Montgomery, 1998).

Courses in the Science and Pseudoscience of Psychology: A Model Syllabus

Largely as a consequence of the renewed interest in the problem of pseudoscience and its importance to education in psychology, a number of instructors across the country, including us, have developed undergraduate courses devoted to science and pseudoscience (Jones & Zusne, 1981; McBurney, 1976; B. Singer, 1977; Swords, 1990; Wesp & Montgomery, 1998). Based on our syllabi and those of several faculty members across the country who have taught closely related courses (see note 1), we present a model syllabus for undergraduate psychology courses in science and pseudoscience (see Table 1). Interested readers can find an excellent set of resources for such courses in B. Singer (1977), who described guidelines for teaching courses focusing on the scientific examination of paranormal phenomena.

This article goes beyond previous articles dealing with the teaching of psychology courses in science and pseudoscience in several ways. In contrast to the courses outlined by B. Singer (1977) and by authors who have outlined psychology courses relevant to pseudoscience in the pages of this journal (e.g., Jones & Zusne, 1981; McBurney, 1976; Wesp & Montgomery, 1998), the course described in this article examines potentially pseudoscientific claims in general rather than paranormal claims per se. Both B. Singer's course and the pseudoscience courses described in this journal, for example, did not focus on such issues as lie detection, subliminal persuasion, questionable assessment procedures, invalidated
Among the topics covered in the course, we emphasize the importance of evaluating claims with explicit reference to the criteria of pseudoscience (e.g., falsifiability, self-correction; Class 1), the fallibility of human reasoning processes (e.g., heuristics, biases, logical errors, and fallacies; Classes 2, 3), and the reconstructive nature of memory, false memory syndrome and potentially suggestive memory recovery techniques, alien abduction reports, multiple personality disorder (Acocella, 1998; Blackmore, 1992; Loftus, 1993; Lynn, Lock, Myers, & Payne, 1997; Nash, 1987; Spanos, 1994).

In summary, we recommend beginning the course with a general introduction to science, the scientific method, and the differences between science and pseudoscience (e.g., falsifiability, self-correction; Class 1), followed by a detailed discussion of issues concerning the fallibility of human reasoning processes (e.g., heuristics, biases, logical errors, and fallacies; Classes 2, 3). The remainder of the course focuses on critically examining specific topics in psychology (e.g., parapsychology, astrology, subliminal persuasion) that illustrate pseudoscientific claims and methods of inquiry. When exposing students to these topics, we recommend making particular efforts to (a) encourage students to maintain an open mind toward the claims in question, but to insist on high standards of evidence before accepting them and (b) evaluate these claims with explicit reference to the issues discussed at the beginning of the course. With respect to the latter, we place particular emphasis on the differences between science and pseudoscience, the ways in which otherwise rational individuals can be led to adopt erroneous beliefs, and logical errors and fallacies.

Although we based this syllabus on a 12-week course that meets once a week for 3 hr, it can be readily modified for courses with diverse schedules. In addition, we have provided (in parentheses following each topic) three to six representative articles, chapters, or both, for each class. These suggested articles illustrate only the types of readings that instructors may wish to assign in conjunction with each topic and do not constitute an exhaustive set of recommended readings. Suggested primary and supplemental texts, which we have generally used in conjunction with a number of the articles listed in the model syllabus, appear in the following section. In addition, a list of more than 50 suggested term-paper and presentation topics is available from the first author on request. We should emphasize that we do not view this model syllabus as set in stone. Instead, we offer it as a flexible template from which instructors can tailor their own course syllabi and readings.

Instructors can also tailor this model syllabus for many content courses in the undergraduate psychology curriculum. For example, one of the authors of this article (Jeffrey M. Loftus) teaches a psychology research methods course in which he contrasts scientific methodologies with pseudoscientific practices in terms of analytical reasoning and critical thinking, the importance of open-minded skepticism, the role of falsifiability, and the methods by which claims are promoted to the general public. This course also incorporates student-run laboratory projects that involve descriptive research on the measurement of paranormal beliefs and replications of experimental research on pseudoscientific practices, such as the use of subliminal self-help tapes.
Useful Primary and Supplemental Texts for Psychology Courses in Science and Pseudoscience

We list 15 primary and 8 supplemental texts that we believe to be useful for the teaching of psychology courses in science and pseudoscience, along with a brief description of the content of each text. In selecting these texts, we have focused primarily on sources published within the past 15 years to ensure coverage of recent topics. We recommend using one or more of these texts (particularly the primary texts) in conjunction with articles, book chapters, or both (see representative references in the previous section) to provide students with an overarching framework for the principal issues addressed in the course. Because Marek et al. (1998) recently evaluated general critical thinking texts in this journal, we examine only primary texts and supplements that focus largely or exclusively on pseudoscientific claims. The recent *Teaching of Psychology* review by Marek et al. (1998) featured only four of these texts: Gray (1991), Gilovich (1991), Schick and Vaughn (1999), and Stanovich (1998). Following the description of each text, we list in parentheses the lecture(s) from the model syllabus in which instructors can most profitably use each text.

**Primary Texts**

Cardena, Lynn, and Krippner (2000) is an edited volume containing a number of interesting and well-researched chapters on a variety of anomalous psychological phenomena, including lucid dreaming, synesthesias, hallucinations, out-of-body experiences, alien abduction memories, and past life memories. Although several of the chapters (e.g., Targ's chapter on ESP) reflect a less skeptical approach to the subject matter than that found in most other texts reviewed here, this book offers a refreshing and open-minded overview of scientific research on many mysterious and purported paranormal experiences (Classes 4, 5, 8, 11, 12).

Carey (1998) provides a succinct and readable introduction to the scientific method and basic principles of scientific thinking, including the use of proper controls and the differences between correlation and causation. The final two chapters ("Extraordinary Claims and Anecdotal Evidence" and "Fallacies in the Name of Science"), which comprise approximately one third of the book, contain entertaining discussions, examples, and exercises focusing on pseudoscientific claims (e.g., graphology, backwards subliminal messages) and their differences from scientific claims (Classes 1–3).

Della Salla (1999) is a good although slightly uneven edited volume that contains interesting chapters focused on myths concerning brain functioning. The chapters on the "People Only Use 10% of Their Brains Myth" and brain tuners (both by Beyerstein) are highly informative and engaging, as are the chapters on near-death experiences, repetition and memory, and myths concerning the differences between left and right hemisphere functioning (Classes 4, 8, 11, 12).

Friedlander (1995) focuses on pathological science and its characteristics. Although much of the book deals with questionable claims in physics and chemistry (e.g., cold fusion, polywater), the volume devotes considerable space to astrology, ESP, and other issues relevant to psychology students. Although this book is more advanced than most other texts on pseudoscience, we recommend it for instructors wishing to provide students with a broad perspective on the nature of pathological science in general (Classes 1, 5, 6).

Gilovich (1991) is a superb and highly readable introduction to social cognitive heuristics and biases, their impact on everyday thinking, and their implications for beliefs in ESP, alternative medicine, and other questionable claims. Gilovich draws heavily on examples from sports, interpersonal relationships, and other issues of interest to undergraduates (Classes 2–5).

Gray (1991) provides a succinct, well-written, and highly accessible introduction to 16 common logical fallacies (e.g., genetic, bandwagon, appeal to ignorance), the differences between science and pseudoscience, and inferential and empirical issues relevant to various purported paranormal phenomena (e.g., ESP, the Bermuda Triangle; Classes 1–5).

Hines (1988) is a classic, although now slightly dated, review of the scientific evidence pertinent to a broad spectrum of paranormal and other extraordinary claims, including spiritualism, ESP, psychoanalysis, astrology, the lunar lunacy effect, biorhythms, and faith healing. Chapter 12 ("Current Trends in Pseudoscience") features brief summaries of 15 other dubious scientific assertions, such as dowsing, firewalking, and Kirlian photography (Classes 1–11).

Neher (1990) provides readers with an open-minded but appropriately skeptical and critical introduction to paranormal, mystical, and occult experiences. It contains good analyses of ESP, psychic healing, out-of-body experiences, transcendental meditation, possession phenomena, ghosts and apparitions, astrology, Pyramid Power, and other extraordinary and unusual claims. Chapter 2 features a good discussion of how humans create psychological meaning even in its objective absence (Classes 2, 4–6, 8, 10–12).

Piattelli-Palmarini (1994) is a solid introduction to cognitive illusions, heuristics, and biases and their implications for odd and erroneous beliefs. The author covers much of the same ground as Gilovich (1991), but this book is somewhat more advanced and difficult (Classes 2, 3).

Sagan (1995a) is the penultimate book by the late Cornell astrophysicist, who spent much of his career educating the public about both the wonders of science and the dangers of pseudoscience. It contains detailed discussions of UFO abduction reports, false memories, and the value of scientific education and critical thinking. Chapter 12 ("The Fine Art of Baloney Detection") is a must for science and pseudoscience courses. Overall, this book is an eloquent and impassioned manifesto for clear thinking (Classes 1, 3, 6, 12).
Schick and Vaughn (1999) provides a thoughtful and interesting introduction to issues in philosophy of science (e.g., relativism) and social cognition (e.g., selective attention) relevant to evaluating pseudoscientific and extraordinary claims. The text is sprinkled liberally with numerous interesting examples drawn from the annals of the odd and unusual. It is slightly more advanced than similar texts (e.g., Gray, 1991), but somewhat more relevant to pseudoscience per se and an countr with respect to recent issues in science and pseudoscience (Classes 1–5, 12).

Shermer (1997) provides an excellent and highly readable discussion of the problem of pseudoscience and pseudohistory in modern America. Among the featured topics are perceptive and engaging analyses of ESP, astrology, creationism, Holocaust revisionism, and other unusual belief systems (Classes 1–5, 12).

Stanovich (1998) is a standard text in critical thinking courses in psychology, although it is also an outstanding introduction to philosophy of science, the differences between science and pseudoscience, and common misconceptions regarding psychology. The chapters on placebo effects, Clever Hans, and probabilistic reasoning are especially relevant to courses in the science and pseudoscience of psychology. On balance, this book is a delight to read (Classes 1–5, 10, 11).

Vyse (1997) was the recipient of the 1998 William James Book Award from the American Psychological Association. This book offers the best available general overview of superstitious thinking and its psychological and sociological origins. In addition, it contains good discussions of the sources of paranormal beliefs in general (Classes 2–5).

White (1999) consists of well-written, engaging, and highly readable critiques of unusual beliefs in many areas of science, including psychology. Like Friedlander (1995), this book is especially well suited for general courses on pathological science. The chapters include critical discussions of telepathy, remote viewing, faith healing, pain control, near-death experiences, voodoo, cults, and alien abductions (Classes 4, 5, 8, 12).

**Supplemental Texts**

Frazier (1986) is a superb anthology of *Skeptical Inquirer* articles featuring discussions of parapsychology, misperceptions and illusions, palm reading, astrology, UFOlogy, creationism, and other topics (Classes 1–7, 10).

Frazier (1991) is another anthology of *Skeptical Inquirer* articles consisting of critical discussions of alien abduction reports, past-life hypnotic regression, alpha consciousness, graphology (handwriting analysis), recent ESP research, spontaneous human combustion, astrology, and chiropractic. The articles by Hyman and others on critical thinking skills are particularly instructive (Classes 1–7, 10, 12).

Frazier (1998) is a superb edited compilation of articles dealing with such topics as polygraphy, self-report honesty tests, parapsychology, near-death experiences, lucid dreaming, facilitated communication, satanic cults, and recovered memories. The opening article by Sagan, “Wonder and Skepticism,” is a superb and inspiring introduction to scientific skepticism (Classes 1–5, 8–12).

Gardner (1991) is a delightful and thought-provoking collection of essays focusing on parapsychology. Among the topics addressed are the exploits of Israeli “psychic” Uri Geller, psychic surgery, psychokinesis, the reincarnation claims of Shirley MacLaine, the peculiar relationship between Sigmund Freud and Wilhelm Fleiss, and the importance of debunking in science (Classes 1, 4, 5).

Gardner (1992) is another highly informative and engaging anthology of articles on pseudoscience. It includes commentaries on ESP, levitation, channeling, Wilhelm Reich’s orgone therapy, glossolalia (speaking in tongues), and the role of astrology in the Reagan administration. Gardner’s discussions of relativism and realism in science are especially provocative (Classes 1, 4–6, 11).

Nickell, Karr, and Genoni (1998) is a succinct edited volume featuring articles on UFO hoaxes, firewalking, hypnotic age regression, and the “cold reading” techniques used by palm readers, astrologers, and crystal ball readers, among other paranormal topics (Classes 1, 4, 6, 7, 11).

Tavers (1995) is an engaging collection of short opinion pieces on a wide variety of controversial psychological topics, including astrology, psychoanalysis, illusory correlation, codependency, and premenstrual syndrome. Instructors who wish to generate lively discussion and debate in small seminars will find this brief book especially useful (Classes 2, 3, 6, 11).

Youngson (1998) is an entertaining collection of succinct vignettes regarding classic errors in science, such as Lamarck’s theory of the inheritance of acquired characteristics, N-rays, cold fusion, and the use of magnets to cure medical ailments. One chapter focuses on erroneous or overblown claims in psychology, including phrenology, psychoanalysis, illusory correlation, and the Rorschach Inkblot Test. This book also contains a brief but useful discussion of “How to Detect Pseudoscience” (Classes 1, 10–12).

**Useful Videos for Psychology Courses in Science and Pseudoscience**

A number of videos, many of them commercially available, offer excellent illustrations and demonstrations of a variety of issues relevant to pseudoscientific psychology and the paranormal. We have found such videos to be valuable supplements to the readings recommended in the previous section and to be helpful springboards for class discussion and debate. In this section, we present 11 video programs that we recommend highly for instructors teaching psychology courses in science and pseudoscience. As in the previous section, we list the classes on the model syllabus most relevant to each video in parentheses following each description.

Beyond Science (Scientific American Frontiers; Huntley & Angier, 1997–1998) is a superb show, hosted by Alan Alda, that features critical discussions of water dowsing and other apparent psychic powers; graphology; the purported crash of a flying saucer in Roswell, New Mexico; and other paranormal claims. It is an excellent vehicle for generating discussion in the initial classes of the course (Classes 1–6, 9, 11).

Case of the Bermuda Triangle (Nova; Massey, 1976) is somewhat dated, although it provides excellent illustrations of pseudoscientific thinking and the systematic debunking of
a paranormal claim. It is useful for generating class discussions of social cognitive errors and biases, including the use of “variable windows” (for a discussion, see Gilovich, 1991), confirmatory bias, and illusory correlation (Classes 2, 3).

*Divided Memories, Parts 1 and 2* (Frontline; Bikel, 1995) is a powerful and disturbing two-part program focusing on the scientific evidence regarding the existence of recovered memories of sexual abuse and on the impact of suggestive therapeutic procedures, such as hypnotic age regression and past-life therapy, on clients. It offers a first-rate psychological and sociological analysis of the recovered memory controversy (Class 12).

*Exploring the Unknown* (Fox Family Channel; Kropnick & Schreiber, 1999–2000) is a multipart program featuring segments hosted by Consulting Producer Michael Shermer, editor of *Skeptic* magazine. Students benefit from balanced but appropriately skeptical segments on cold reading, astrology, remote viewing, polygraph testing, and related topics (Classes 4–6, 9–12).

*Kidnapped by UFOs* (Nova; DiFanni, 1996) offers a good introduction to the alien abduction phenomenon and features thoughtful analyses by Elizabeth Loftus, the late Carl Sagan, and other experts. Although this program does not provide as extensive a discussion of the psychology of memory as might be desired, it illustrates a number of important issues concerning the reconstructive nature of memory and the perils of suggestive questioning (Class 12).

*Liar* (A & E Voyages and BBC TV; Claxton, 1996) is an entertaining program offering an excellent introduction to scientific controversies regarding the polygraph test and other purported lie-detection methods. Also featured are intriguing demonstrations of autistic children’s apparent incapacity for genuine deception and a possible instance of lying in chimpanzees (Class 9).

*Multiple Personality Disorder* (The Fifth Estate, 1993) provides a first-rate discussion and critical analysis of the controversial issues surrounding the diagnosis of multiple personality disorder (now referred to as dissociative identity disorder). Experts interviewed on this program place particular emphasis on the potentially iatrogenic effects of suggestive questioning, hypnosis, and other questionable therapeutic practices. They also call into question much of the commonly held wisdom regarding multiple personality disorder and its causes (Class 12).

*The Power of Belief* (ABC News; Ellis, Golden, & Matthews, 1998) is a highly engaging and informative video narrated by John Stossel. Along with Nova’s *The Secrets of the Psychics* and Scientific American Frontier’s *Beyond Science*, this program is the best general introduction to issues in pseudoscience and the paranormal available on video. Among the issues investigated are ESP, astrology and tarot card reading, firewalking, voodoo, faith healing, and the placebo effect (Classes 1–6, 9, and 11).

*Prisoners of Silence* (Frontline; Dalfreman, 1993) is an outstanding and provocative video that provides a critical evaluation of facilitated communication, a technique widely claimed in the 1980s and early 1990s to permit autistic children to communicate effectively with others. This program almost invariably has a profound impact on students. We regard this video as a must for instructors wishing to instill critical thinking in their students concerning psychotherapy and psychotherapy research design (Classes 1, 2, 11).

*Secrets of the Psychics* (Nova; Charlson, 1993) is a superb and highly entertaining video featuring the exploits of magician James (“The Amazing”) Randi, who for the last several decades has been a tireless debunker of pseudoscientific claims. This program highlights critical evaluations of ESP, astrology, faith healing, palm reading, and other ostensible paranormal phenomena. We highly recommend this video as an introduction to issues concerning pseudoscience in general and parapsychology in particular (Classes 1–6, 10).

*The Skeptic’s Guide to the Paranormal* (Discovery Channel; McCabe & Burns, 1999) provides first-rate critical discussions and demonstrations of several purported paranormal phenomena. The segments feature James Randi debunking psychics and revealing some of the secrets underlying simple magic tricks, aircraft and space experts providing mundane explanations for UFO reports, and photographic experts explaining otherwise mysterious images of ghosts and apparitions. In addition, this program contains a thoughtful discussion of placebo effects and their role in the apparent effects of therapeutic touch and other dubious medical treatments (Classes 2–6, 11).

**Useful Web Sites for Psychology Courses in Science and Pseudoscience**

Over the past several years, a large number of Web sites providing critical evaluations of potentially pseudoscientific claims have appeared. These sites vary widely in quality, with some offering impartial and research-based analyses of these claims and others offering little more than subjective opinions. Many of the high-quality sites are excellent resources for students researching topics in pseudoscience, as well as useful jumping-off points for class discussion and debate. Table 2 provides a list of 12 Web sites that we recommend to both instructors and students.

**Will Teaching Courses in Science and Pseudoscience Make a Difference?**

One final but crucial question that must be addressed is whether psychology courses in science and pseudoscience accomplish the goal of facilitating students’ critical thinking regarding paranormal and other extraordinary claims. One source of evidence derives from student evaluations of teaching (SETs). Wesp and Montgomery (1998) reported that SETs following a course entitled Experimental Investigation of the Paranormal were generally quite positive. We have similarly found that students evaluate courses in science and pseudoscience very positively. One of the authors of this article (Scott O. Lilienfeld), for example, recently taught an advanced undergraduate seminar entitled Science and Pseudoscience in Psychology: Thinking Critically About Human Behavior for three semesters at Emory University. Across all three semesters (Ns = 9, 8, and 13, respectively), the mean responses to the question “How much did you learn from the course?” on a scale ranging from 1 (nothing) to 9 (a lot) were 8.78 (SD = 0.66), 8.63 (SD = 0.52), and 8.23 (SD
The critical thinking literature in education has reported mixed results concerning the relation between psychology courses in science and pseudoscience and critical thinking in general, although further controlled studies of this issue are warranted. We should point out, however, that the critical thinking literature in education has reported mixed results concerning the relation between belief in the paranormal and critical thinking in general (e.g., Morgan & Morgan, 1998). Thus, the extent to which psychology courses in science and pseudoscience generalize to critical thinking skills in other domains remains a critical but largely neglected topic of investigation. In addition, the
References

Asterisked references are recommended readings listed in the model syllabus (see Table 1).


Bikel, O. (Producer). (1995). Divided memories, Parts 1 and 2 (Frontline). Boston: WGBH. (Available from WGBH, 125 Western Avenue, Boston, MA 02134)


Notes

1. We thank Ray Hyman (University of Oregon), Michael Kane (Georgia State University), Stuart Vyse (Connecticut College), Chris Wetzel (Rhodes College), Anthony Pratkanis (University of California, Santa Cruz), Robyn Dawes (Carnegie Mellon University), and Donald Jensen (University of Nebraska) for sharing their syllabi, suggested readings, and other course materials.

2. Please send correspondence to Scott O. Lilienfeld, Department of Psychology, Room 206, Emory University, Atlanta, GA 30322; e-mail: slilien@emory.edu.
Helping Students Read Reports of Empirical Research

Connie K. Varnhagen  
University of Alberta

Nancy Digdon  
Grant MacEwan College

“Reading the Research” is a Web-based interactive learning module designed to help students learn to comprehend and critically evaluate reports of empirical research. The module can be adapted for use with any report of empirical research. Students work through the module, take self-tests for factual knowledge, and discuss critical-thinking questions in an asynchronous discussion group. We conducted a formative evaluation with upper level experimental psychology students and a summative evaluation with introductory students. Students had positive attitudes toward the module and believed it improved their learning. Increased use of the module was also correlated with better learning of the report of empirical research.

An essential goal of the undergraduate psychology curriculum is learning to read and think critically about research (Baum et al., 1993; Brewer et al., 1993; Halpern, 1998; McGovern, 1993; Miller & Gentile, 1998). Students, especially those intent on graduate studies, must gain experience reading reports of empirical research. However, they often find the style and level of writing considerably more difficult than textbooks. Many students are intimidated and overwhelmed by reports of empirical research. All too often, they simply read the arguments in the introduction and discussion sections and fail to critically evaluate the hypotheses, methods, results, or the conclusions sections. The purpose of the Reading the Research Web-based interactive module is to assist students in learning to comprehend and read critically empirical reports.

Description of Reading the Research

Reading the Research provides a summary of a research report laid out in American Psychological Association (APA) style sections. The module models the format of a report, helping students become familiar with the structure and conventions specified by the APA (2001).

The first part of the module describes the genre of the report of empirical research and presents a set of questions a critical reader considers while reading a report. Guides for critically reading reports of empirical research describe a strategy of asking questions while reading different parts of the report (e.g., APA, 2001; Anisfeld, 1987; Chamberlain & Burrough, 1985; Meltzoff, 1998). We adapted questions from these guides and categorized them as factual or critical-thinking questions. The questions we used in Reading the Research appear in Table 1. Factual questions are important because they assess whether the student knows the specifics of the research well enough to evaluate it. Questions that require inferences and critical thinking are open ended to encourage students to think creatively and critically about what they are reading.

The second part of Reading the Research is a summary of the report of empirical research and is accompanied by the factual and critical-thinking questions a critical reader should consider while reading the different sections of a report of empirical research. Students work through the summary and address the questions while reading the report. This part of Reading the Research comes as a template that an instructor can customize for use with a particular report. The instructor modifies the template by editing the summary and question pages in a word processing program and then publishes it on his or her Web site.

The template has a separate Web page for each section of an APA-style report of empirical research. For each of the sections, the instructor inserts information that is specific to the empirical report that students are trying to comprehend and evaluate. An instructor can add more or less information, depending on the level of students and on the difficulty of the report. The instructor can include a summary of the key points described in the section, background information that provides the broader context of the research, methodological or ethical issues, or an elaboration of difficult concepts. Furthermore, an instructor can include links to Web resources related to the research (e.g., a page describing different types of research designs or a glossary of new terms) or to outside information available on the Web (e.g., information about the authors or the journals, links to similar research, a picture of apparatus used in the research, a news report on the research). Links enable the instructor to bring in the wealth of information available on the Web.

1 An example Reading the Research module is available at http://www.psych.ualberta.ca/~varn/Kenrick/Reading.htm
Table 1. Factual and Critical-Thinking Questions Included in Reading the Research

<table>
<thead>
<tr>
<th>Factual Questions</th>
<th>Critical-Thinking Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who did the research?</td>
<td>Where did these research questions come from?</td>
</tr>
<tr>
<td>Where was it published?</td>
<td>Is the research important? Why or why not?</td>
</tr>
<tr>
<td>What are the research questions?</td>
<td>Are the research participants appropriate for the study?</td>
</tr>
<tr>
<td>Who are the participants in the study?</td>
<td>Is the research design appropriate for the research question(s)?</td>
</tr>
<tr>
<td>What is the research design?</td>
<td>Are the measures appropriate for addressing the research question(s)?</td>
</tr>
<tr>
<td>What are the measures?</td>
<td>What ethical considerations are important to address? Are they all addressed in the article?</td>
</tr>
<tr>
<td>What are the main results of the study?</td>
<td>Can the results be used to answer the research question(s)?</td>
</tr>
<tr>
<td>What conclusions do the researchers draw from the results?</td>
<td>Can the results be generalized beyond the context of the study?</td>
</tr>
<tr>
<td></td>
<td>Are the conclusions important? Why or why not?</td>
</tr>
</tbody>
</table>

Reading the Research is an interactive module. Each of the summary pages has clickable icons linked to multiple-choice test questions or to online discussion groups. The template includes all the programming required to format multiple-choice questions and to provide students with performance feedback. An instructor simply types in the relevant content. Use of the online discussion groups requires a special program; we use Discus (Paulisse & Polik, 2000) or WebCT (http://www.webct.com/). Typically, the instructor begins the discussion by posting a question or comment. Students respond to the instructor’s posting or to postings of other students. Online discussion groups are especially useful for attaining participation from students in large classes. Furthermore, for classes with graded participation, the online discussion may facilitate evaluation because a copy of student contributions is preserved. Moreover, online discussion enables more reflective discussion in that students have the opportunity to ponder questions and comments before responding to them (Varnhagen, Drake, & Finley, 1997). They do not have to react right away, as they often must do during an in-class discussion. In addition, all students in a large class can engage in discussion when grouped into small asynchronous discussion groups, whereas they may not be willing or able to in a large lecture hall.

Formative Evaluation in Experimental Psychology

We conducted a formative evaluation of Reading the Research in a second-year experimental psychology class. As part of a laboratory exercise, students completed a module, written to accompany an article on day care (Broberg, Wessels, Lamb, & Hwang, 1997), and then evaluated the module; we did not grade participants on their performance. The goals of this formative evaluation study were to determine students’ perceptions of the usefulness of Reading the Research for learning how to read reports of empirical research and to evaluate whether students perceived that Reading the Research helped them learn about the format and content of reports of empirical research.

Method

Participants. Participants were 5 male and 11 female college students enrolled in a second-year experimental psychology course. The students were somewhat experienced at using computers. When asked about their computer experience, the modal response was that they were “moderately comfortable” working on computers and had “fair” computer skills. Only 1 student reported no experience with computers. They reported moderate computer access (63% had a computer at home), but minimal Internet access (26% had access at home). As part of course requirements, students read APA-style reports and used them as reference material in papers. For most of the students, this was their first exposure to empirical reports.

Materials. The second author created a Reading the Research module with discussion questions for Broberg et al. (1997). We both developed a questionnaire designed to assess students’ perceptions of the module. The questionnaire had 5-point Likert scale items to assess attitude toward the module, yes–no questions about difficulties encountered using the module, and multiple-choice items about students’ preferences for activities associated with reading reports of empirical research. These questions appear in Table 2. We measured student self-ratings of computer skill through several Likert scale questions, including “How comfortable do you feel working on computers?” ranging from 1 (not at all comfortable) to 5 (very comfortable).

Procedure. Students met as a group in a computer lab for a single 2-hr session to work through the Reading the Research module. They had been instructed to read the report associated with the module (i.e., Broberg et al., 1997) before attending the lab session and to bring it to the session. Students worked through the module at their own pace, and contributed to the online discussions as part of their lab assignment. Because anyone on the Web could read the discussion, students posted their comments using first names only or a pseudonym. At the end of the lab session, students completed the questionnaire, which evaluated their use of the module.

Results and Discussion

Table 2 presents the students’ modal attitudes toward the module. The students expressed positive attitudes. No student rated Reading the Research as being “not at all enjoyable.” Students rated the Reading the Research module as the most helpful activity for comprehending empirical reports, compared with reading the report, discussing the report in class, or...
being tested on the report. They reported that they thought they had learned something about the specific research and about APA-style reports in general. They rated reading the actual report as least helpful for comprehending it. This finding is consistent with anecdotal evidence suggesting that students find reports of empirical research dry and difficult, sometimes so much so that they are not motivated to read them thoroughly enough to be able to think critically about them (Baum et al., 1993; Brewer et al., 1993; McGovern, 1993; Meltzoff, 1998; Wade, 1997). Reading the Research seems to be one tool that instructors can use to make the empirical literature more accessible to undergraduates.

Some attitudes toward the module were correlated with self-reports of computer skills. Students who rated themselves as being more comfortable in using computers had a more positive overall impression of the module than did students who rated themselves as less comfortable, $r(14) = .76$, $p < .05$. However, students who rated themselves as being less comfortable using computers enjoyed the module more, $r(14) = .49$, $p < .05$. This finding may have been a contrast effect; students with poorer computer skills may have been surprised that they could navigate the module successfully and therefore found the experience more enjoyable than anticipated. The other correlations between comfort and attitude were small and not statistically significant. These results indicate that even novice computer users have a positive attitude toward using a Web-based supplement in their learning.

This study demonstrated that students in a second-year experimental psychology course found the module useful and enjoyable. In the next study, we examined how students used the module in a course setting and how their use related to learning.

Evaluation in Introductory Psychology

We integrated Reading the Research into “intro.psych,” a technologically enhanced introductory psychology course (Varnhagen, 1999; Varnhagen, Winship, & Apedoe, 2000). Students in intro.psych attend a keynote lecture on each topic in the introductory psychology curriculum and then complete Web activities designed to enhance their understanding of these different topics. An important component of the Web activities is the students’ asynchronous discussion. Students participate in discussion groups of three to five students, and our course technology (Heth, 1999) allows students to link directly to a specific thread in their discussion group (Paulisse & Polik, 2000).

Students completed two Reading the Research modules, one accompanying research on adolescent dating (Kenrick, Gabrielidis, Keefe, & Cornelius, 1996) and one on preparing for examinations (Balch, 1998). These reports of empirical research, and the accompanying Reading the Research modules, were a part of course requirements. Students were graded on their asynchronous discussion of the reports. Questions relating to comprehension of the articles and understanding of the structure of reports of empirical research appeared on the multiple-choice midterm and final examinations.

We examined students’ performance measures as a function of their use of the modules as indicated by server log files. If Reading the Research and the accompanying discussion helped students learn about and critically evaluate empirical research, we would expect to see relationships between use of the module and performance based on critically examining and learning about the research from the module.

Method

Participants. Participants were 46 male and 50 female students enrolled in intro.psych. Of these, 66 students were in their first year of university, 17 were in their second year, 7 were in their third or higher year, and 6 were unclassified students. All students participated in the evaluation as a part of course credit for research participation.

According to an online survey conducted the first week of the term, few students were novice computer or Web users: 60% rated themselves as “comfortable” or “very comfortable” with using computers, 65% rated their Web browsing skill level as “excellent” or “good,” and 69% indicated they had home access to the Web. The course catalogue listed the course as having an Internet component; thus self-selection may have led to students having better-than-average computer skills and access.

Procedure. Students completed the Reading the Research modules as a part of their regular class work. They completed the first module, accompanying Kenrick et al. (1996), approximately one third of the way through the course, and they completed the second module, accompanying Balch (1998), at the end of the term. Students were quite accustomed to using the Web components of the course by these

<table>
<thead>
<tr>
<th>Table 2. Modal Responses to the Questionnaire Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Questionnaire Item</td>
</tr>
<tr>
<td>-------------------</td>
</tr>
<tr>
<td>What is your overall impression of Reading the Research?</td>
</tr>
<tr>
<td>How interesting did you find Reading the Research?</td>
</tr>
<tr>
<td>How enjoyable did you find Reading the Research?</td>
</tr>
<tr>
<td>How worthwhile did you find Reading the Research?</td>
</tr>
<tr>
<td>Do you think this module should be a part of the class?</td>
</tr>
<tr>
<td>What did you do that helped you understand the empirical research article the most?</td>
</tr>
<tr>
<td>What did you do that helped you understand the empirical research article the least?</td>
</tr>
<tr>
<td>Do you feel you learned something about the research?</td>
</tr>
<tr>
<td>Do you feel you learned something about the different parts of a report of empirical research?</td>
</tr>
</tbody>
</table>
times. After a keynote lecture introducing the concepts, students had 1 week to complete the assigned Reading the Research module. We asked students to complete an online questionnaire similar to the one used in the formative evaluation following completion of the module. One week after doing the module, students completed either a midterm or final examination, which included seven multiple-choice questions about the associated report of empirical research.

As part of a larger research project on student learning from the Internet (Varnhagen et al., 2000), students also completed the Cornell Critical Thinking Test (CCTT; Ennis & Millman, 1985), an online questionnaire regarding their computer skills during the first week of the term, and an online questionnaire regarding their impressions of the course during the last week of the term.

**Results and Discussion**

Fewer than 20% of the students completed the online evaluations for the Reading the Research modules. Although we do not report the results here due to the small return rate, the students’ impressions were comparable to those obtained in the formative evaluation.

We examined student computer log files to determine how students used Reading the Research. We examined how many and what pages students accessed and how many times they posted to their discussion group. Table 3 shows the statistics for number of sessions, total number of pages accessed, number of pages accessed by type of page (the introductory pages on how to read reports of empirical research, the summary pages, and the multiple-choice factual test pages), number of discussion posts, percentage mark for the discussions, and percentage score on the exam questions related to the module.

Most of the students worked with Reading the Research in more than one session. Students generally worked through the module in a linear fashion during the first or second session and then revisited selected pages, such as the introduction, results, and discussion summary pages, in additional sessions. A few students appear to have bypassed the summary pages altogether, jumping straight to the graded discussions.

We analyzed differences between each of the measures shown in Table 3 with two-tailed paired t tests. Few differences were obtained, indicating consistency in use of the module and performance on the assessments for the two reports. Students accessed more of the summary Web pages when working through the Kenrick et al. (1996) module than they did while working through the Balch (1998) module, t(93) = 1.98, p = .051, ω² = .03. This difference is not surprising because some of the Web pages—those that describe the genre of empirical reports—are the same in different modules. Once students have accessed them from the first module, they may not have a need to revisit them in subsequent modules. Students also accessed more of the multiple-choice test pages for the Kenrick et al. module than for the Balch module, t(93) = 3.70, p < .01, ω² = .05, for the analysis of weighted means. This difference is particularly interesting because the multiple-choice midterm exam questions came directly from the module; possibly students did not realize this fact and decided they did not need to check their factual knowledge for the Balch article. Finally, students performed better on the multiple-choice exam questions related to the Balch article than they did on the exam questions related to the Kenrick et al. article, t(92) = −3.06, p < .01, ω² = .07. The remaining comparisons were not statistically significant.

We used two-tailed correlational analyses to examine the relation between use of the different components of Reading the Research and learning, as measured by exam performance. These correlations appear in Table 4. Only number of discussion posts made correlated with exam performance for the Kenrick et al. (1996) module, but all components of the module correlated with performance for the Balch (1998) module. Given that we found few mean differences in students’ use of the modules from the first reading to the second, these differences in correlations may indicate that the students became better able to use Reading the Research in their learning on their second set of experiences with the module.

We also analyzed correlations between other variables and learning. Critical-thinking ability, as measured by the CCTT administered at the beginning of the term, was not correlated with exam performance. In addition, preference for Internet-based courses versus traditional lecture-based courses did not correlate with performance. This preference variable consisted of the student’s response to the question, “What do you think of the way this course was taught [with the Internet component] compared to a more traditional class lecture format?” Student responses to this variable were bimodal: 46% responded that they either somewhat or much preferred the Internet, 52% responded that they somewhat or much preferred a traditional format, and 3% responded that they had no preference. The small and nonsignificant correlation between preference and learning indicates that students can learn from technology-based courses even if they do not particularly like the instructional format.

We used stepwise multiple regression analyses to determine which components of the Reading the Research modules contributed significant independent variance to learning. The predictor variables included the number of

---

**Table 3. Means and Standard Deviations for Sessions Completed, Pages Accessed, and Exam Performance**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sessions</td>
<td>5.1 5.7</td>
<td>4.9 5.6</td>
</tr>
<tr>
<td>Total pages accessed</td>
<td>24.6 19.9</td>
<td>17.5 19.8</td>
</tr>
<tr>
<td>How-to-read pages</td>
<td>4.8 4.3</td>
<td>4.4 6.0</td>
</tr>
<tr>
<td>Summary pages</td>
<td>12.3 11.1</td>
<td>10.5 11.5</td>
</tr>
<tr>
<td>Multiple-choice pages</td>
<td>6.9 7.4</td>
<td>2.5 4.1</td>
</tr>
<tr>
<td>Discussion posts</td>
<td>3.0 2.7</td>
<td>3.8 3.3</td>
</tr>
<tr>
<td>Discussion grade</td>
<td>67.7 29.2</td>
<td>64.7 36.5</td>
</tr>
<tr>
<td>Examination grade</td>
<td>52.2 19.2</td>
<td>60.3 18.9</td>
</tr>
</tbody>
</table>

*Based on 24 total Web pages in the Kenrick et al. module and 18 total Web pages in the Balch module. There were 2 how-to-read pages for each module, 7 summary pages for each module, and 15 multiple-choice pages for Kenrick et al. and 9 multiple-choice pages for Balch. *Given in percentages.
Table 4. Correlations With Exam Performance and Sessions Completed, Pages Accessed, Discussion Posts, Critical Thinking, and Course Preference

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sessions</td>
<td>–.07</td>
<td>.20</td>
</tr>
<tr>
<td>Total pages accessed</td>
<td>.15</td>
<td>.29*</td>
</tr>
<tr>
<td>How-to-read pages</td>
<td>.16</td>
<td>.22*</td>
</tr>
<tr>
<td>Summary pages</td>
<td>.12</td>
<td>.29*</td>
</tr>
<tr>
<td>Multiple-choice pages</td>
<td>.13</td>
<td>.25*</td>
</tr>
<tr>
<td>Discussion posts</td>
<td>.28*</td>
<td>.28*</td>
</tr>
<tr>
<td>Discussion mark</td>
<td>.08</td>
<td>.20</td>
</tr>
<tr>
<td>CCTT score</td>
<td>.17</td>
<td>.28</td>
</tr>
<tr>
<td>Course preference</td>
<td>.11</td>
<td>.18</td>
</tr>
</tbody>
</table>

Note. CCTT = Cornell Critical Thinking Test. *p < .05.

how-to-read pages accessed, the number of summary pages accessed, the number of multiple-choice test pages accessed, and the number of discussion posts made. The predicted variable was examination performance. Consistent with the correlational analyses, number of discussion posts was the only significant predictor of exam performance for the Kenrick et al. (1996) module, accounting for 7% of the variance, F(1, 93) = 7.88, p < .01, ω² = .07. Although all components of the Reading the Research module correlated significantly with exam performance for the Balch (1998) article, only accessing the summary pages contributed a significant amount of independent variance (7%) to the exam score, F(1, 92) = 8.21, p < .01, ω² = .07.

General Discussion

Experimental psychology students reported that they believed Reading the Research enhanced their comprehension of a report of empirical research. Introductory psychology students, many of whom had never been exposed to the genre of a scientific report, were able to use the module to aid their learning. Exam performance and relations between use of the module and examination performance increased from their first to their second set of experiences with Reading the Research.

Many students heartily embrace new forms of instructional technology, only to abandon them after the novelty wears off (Clark, 1983). Although two sets of exposures to Reading the Research can hardly be considered overexposure, these results demonstrate that students did use the module in their learning after the initial novelty had worn off.

Providing students with a Web-based summary of a report of empirical research and interactive opportunities such as self-tests and asynchronous discussion thus appears to help students learn both about the difficult genre of the report of empirical research and about the research contained within such a report. Based on the results of our evaluation with introductory psychology students, asynchronous discussion seems to be a particularly important component. This finding adds to previous work (e.g., Varnhagen et al., 1997) demonstrating that Web-based discussion enhances critical-thinking, reading, and writing skills.

We developed Reading the Research as a flexible module to supplement any report of empirical research and by students of differing levels of skill in reading psychological literature. Reading the Research has the potential to assist undergraduate students in acquiring essential critical reading skills and motivate them to read other reports of empirical research.

References


Varnhagen, C. K. (1999, June). Developing and delivering effective instructional technology. Presentation at the meeting of the Society...
TOPICAL ARTICLES

Effects of Instruction in Methodological Reasoning on Information Evaluation

Barry Leshowitz
Kristen Eignor DiCerbo
Morris A. Okun
Arizona State University

In this article we describe an instructional program that focuses on applying causal reasoning and related principles of the scientific method to problems faced in daily life. In a highly interactive classroom setting, the instructor gives students repeated opportunities to apply methodological reasoning to real-world scenarios for the purpose of making informed decisions. In addition to describing the program, we report the findings of a capstone exercise that examined changes in students’ beliefs toward legalization of marijuana after reading persuasive communications. Students who experienced the instructional program exhibited less bias in evaluating information and less attitude polarization than students in a comparison group. We discuss the implications of these findings for developing and evaluating instructional programs in methodological reasoning in psychology.

Over the past 20 years, researchers have demonstrated that human reasoning deviates markedly from what the principles of methodological reasoning prescribe (Stanovich & West, 1998). Research in social cognition has shown, for example, that people’s prior beliefs bias their interpretations of media presentations (Vallone, Ross, & Lepper, 1985), social judgments (Darley & Gross, 1983), impression formation processes (Neuberg, 1994), and assessment of covariation (Chapman & Chapman, 1969). When asked to form and test hypotheses, researchers have shown that people often ask questions that have a greater chance of confirming their theory or belief than of disconfirming it (Darley & Gross, 1983; Snyder & Cantor, 1979; Snyder & Swann, 1978; Wason, 1960).

The first author developed an instructional program in methodological reasoning in an effort to deal directly with these cognitive biases and students’ reliance on scientifically unsupported opinions, beliefs, and hunches. Methodological reasoning involves application of the rules of the scientific method that emphasize posing questions in terms of relationships between variables, formulating alternative hypotheses, testing these hypotheses, collecting data, drawing (causal) inferences, reaching warranted conclusions, and making informed decisions. These “are the kinds of skills that people must have in some measure to live effectively in the world” (Lehman, Lempert, & Nisbett, 1988, p. 441).

The first author incorporated the program in methodological reasoning in his recitation section in research methods for psychology majors and in a course in effective thinking that meets the university’s general studies requirement in critical inquiry (Leshowitz, DiCerbo, & Symington, 1999; Leshowitz, Jenkens, Heaton, & Bough, 1993; Leshowitz & Yoshikawa, 1996). In this article we describe the first author’s program and illustrate how he used a capstone exercise to determine whether the students achieved the goals of the instruction.

Instructional Program in Methodological Reasoning

Subsequently we describe several active-learning, discovery-based techniques used in the recitation section of the research methods in psychology course. The course also included a separate laboratory section that presented several traditional laboratory-based exercises in experimental psychology. In the recitation, students applied principles of methodological reasoning to problems faced by students at school, on the job, and in their personal lives. The course addressed personally meaningful, often highly charged issues, such as the psychological effects of divorce on children, racial profiling, binge drinking, treatment of criminals, and use of placebo control in clinical trials involving terminally ill patients.

The class has no lecture; rather each meeting involves a discussion and analysis of a problem. Additionally, there is no textbook for the course. The instructor integrates the principles of methodological reasoning directly into the analysis of information bearing on the problem. The students prepare for each discussion by reading several articles assigned by the instructor from the popular media. These articles present the kinds of information one often encounters in everyday life. Additionally, the students complete several worksheets that explore various facets of the topic. The lessons often challenge the unquestioned opinions and beliefs of students on current societal issues. The class discussion attempts to make the case for the need to support personal theories with sound evidence.

Sample Lesson

To illustrate how the class might apply principles of methodological reasoning to the analysis of an everyday life situa-
tion, we describe a sample inquiry-based discussion of the effects of divorce on children (readers may request a detailed lesson plan for this topic from the first author). The classroom dialogue concentrated on one article (“The Lasting Wounds of Divorce,” Toufexis, 1989) that appeared in a respected news magazine (Time). The article began as follows: “Many youngsters may still be suffering a decade after breakups. Divorce, everyone agrees, is a traumatic event for children” (p. 61). To support the theory that children of divorce are prone to severe psychological problems, the journalist presented a quote from Judith Wallerstein: “Almost half of children of divorce enter adulthood as worried, under-achieving, self-deprecating and sometimes angry young men and women” (p. 61). To further underscore her conclusion, the journalist presented a vivid description of several children of divorced families. For example, “Ten years after the breakup, Kevin had been in jail three times for beating up a girlfriend, drunken driving and dealing drugs” (p. 61). Furthermore, in the case of Deborah, “Fifteen years later, she is a top student in college, but she has a habit of falling in love with ‘jerks’” (p. 61). Finally, to further strengthen her position on the harmful effects of divorce, the journalist cited a few cursory statistics from a study by Wallerstein and Blakeslee (1988) that investigated 131 children—all of whom had experienced divorce in their families: Three of five youngsters felt rejected by at least one parent, and two thirds of the girls as young adults were unable to make lasting commitments. Recently, Time magazine (Kinn & Pollitt, 2000) presented very similar findings from a controversial follow-up study conducted by Wallerstein and Lewis (1998).

Students critically analyzed the information in the article and reevaluated the basis of their beliefs on the effects of divorce on children, which often were consistent with the conclusions of the media piece, in a reflective dialogue. In this discussion, the first author emphasized a close examination of the evidential basis of the journalist’s thesis. In this effort, he suggested a list of methodological criteria or rules for evaluating empirical evidence, using the following questions: What data were collected in the study presented in the article? How was the study carried out? Do the data show an association between divorce and psychological symptoms? Could other factors besides the divorce of their parents have contributed to the purported behavioral problems? Was the journalist unbiased in the collection of statistics and testimonies of experts? Does the presence of inflammatory language indicate a possible agenda of the journalist? Summarizing the critical analysis, he asked the students: What do you know about this issue and how do you know it?

We found that follow-up activities were useful in solidifying the principles of methodological reasoning introduced in the reflective dialogue. For example, we asked students to redesign the flawed empirical study to meet the criteria for validity. Also, role-playing activities were useful in relating evidence-based decisions to real-life judgments. In the divorce example, we asked students to role-play a subscriber to a national magazine and write a letter to the editor questioning the validity of the divorce article. Asking students to role-play the editor of the magazine who was responsible for accepting or rejecting the article also facilitated the development of multiple perspectives.

In the lesson, students often altered their position dramatically after critically assessing the study’s flawed methodology, the agenda of the journalist, and their own biases. They came to understand that the purported empirical relation between divorce and psychological problems does not come close to meeting the criteria for causality. Without a control group, the students concluded that the association was essentially illusory. Because this revision in their thinking occurred within a context of high personal relevance, students generally showed a sincere willingness to use methodological analysis to evaluate their beliefs and personal theories.

With the class’s increasing demand for “facts,” “scientific evidence,” and “valid information,” students gradually became aware of the impact of their values and biases of others on their thinking. Aware of the effects of emotionally arousing communication that were often deficient in supporting evidence, students began to reevaluate their confidence in the pronouncements of others, especially those appearing in the popular media. The basis of this emerging skepticism was an evidentiary analysis of the relevant information.

Assignments

In the recitation section, in which enrollment may vary from 25 to 80, students completed a plethora of written assignments: in-class and take-home exams (about four), daily homework assignments (brief essays), journal entries (called class reviews), at least three substantial papers about five pages in length, and a cumulative final exam. Students completed about 45 writing assignments, totaling well over 100 pages. Some examples of the real-world-based formats of the assignments included letters to the editor, workplace memora-

Peer Teaching

Some students were reticent about confronting such highly controversial issues in a classroom setting, especially at the outset of the course. To help overcome this problem, the instructor enlisted the assistance of peer tutors in an effort to establish a community of inquiry. Each semester, he selected outstanding undergraduates from previous semesters to serve as peer tutors. Given course credit for their participation in the program, peer tutors provided assistance in the course and received on-the-job training in providing feedback and grading. The instructor graded 20% of peer-tutor graded work throughout the semester and read approximately 50% of assignments to maintain standards. By providing detailed written comments and assessments of assignments, the peer tutors and the instructor helped students explore the assumptions on which they based their personal opinions and beliefs. They asked for
clarification of terms, reasoned support for opinions, acknowledgment of the influence of a personal bias, advantages of the opposing viewpoint, and consideration of values that might affect students’ thinking processes. In this highly supportive learning environment, students realized that the instructor and peer tutors took their ideas, contributions, successes, and frustrations into account.

Capstone Exercise

To provide further insight into the previously described intervention in methodological reasoning, we describe a capstone exercise that gave students an opportunity to bring together the principles of informed decision making. In this exercise, students examined issues relating to the legalization of marijuana. The exercise used the experimental paradigm developed by Lord, Ross, and Lepper (1979) for measurement of cognitive bias and attitude change. First, the students read a brief fictional article that either supported or opposed the legalization of marijuana. The article presented findings from a fictitious, methodologically flawed study. Flaws included the use of small (biased) samples, failure to consider base-rate information, absence of control groups, use of self-selected participants with associated third-variable confounds, and reliance on anecdotal evidence. Students assessed the quality of the article and reported whether they were willing to change their attitude toward legalization of the drug based on the information in the article. The participants then repeated the evaluation for a methodologically flawed article presenting findings that were in opposition to the original article’s findings. The order of the articles was counterbalanced.

Comparison Groups

Sixty-nine students (primarily juniors) enrolled in the first author’s research methods class comprised the trained group. For purposes of comparison, an untrained group of students who were either enrolled in introductory psychology (n = 108) or served as university academic tutors (n = 49) participated in the exercise. The tutors (primarily juniors) received $5 for their 1 hr of participation in the study. Participants in the introductory psychology class partially fulfilled a course requirement by participating in this study. The tutors and the introductory psychology students had not taken the first author’s research methods class.

Because random assignment in this course evaluation was not possible, the presence of possible confounds preclude us from making causal inferences about the effectiveness of the research methods course. For example, the research methods class had taken at least one more psychology class than students in the other groups. These students may therefore have had some previous exposure to methodological reasoning. A close examination of other possible threats to validity revealed that preexisting differences in grade point average (GPA) among the groups probably would not account for possible differences in the outcome of the exercise. Although GPA was unmeasured in this study, university statistics show that students enrolled in research methods and introductory psychology have average GPAs of 3.01 and 2.74, respectively. All of the tutors had GPAs of at least 3.5. To the extent that GPA is related to the evaluation of information, one would expect the tutors to do better than students in the other two courses. Summarizing, the procedure used to select the groups in the exercise minimized the threat to validity of GPA differences but not the threat due to previous exposure to psychology.

A graduate student in psychology who did not know the students explained the exercise and completed distribution and collection of materials. The graduate student indicated to the participants that the study consisted of an investigation of college students’ opinions on various topics. In the debriefing, none of the participants either connected the exercise to the first author or stated the hypotheses addressed in the exercise.

Articles

The students read each article in two installments: The first consisted of a brief statement of the study’s results, whereas the second included a detailed description of the results of the study, researchers’ procedures, several criticisms of the study, and the authors’ rebuttals to those criticisms.

Measures

Initial attitudes. Prior to reading the articles, we assessed students’ initial attitude toward the legalization of marijuana using a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The number of neutral students among the trained and untrained groups was 10 and 39, respectively. In accord with Lord et al.’s (1979) procedure, we did not analyze data from the neutral respondents.

Study quality ratings. To assess the students’ evaluations of the quality of the two studies, after the participants read a detailed description of either the “pro” or “anti” study, the students responded to the following item: “Based on the description of the study, how well or poorly do you think this study was conducted?” Participants responded using a 17-point scale ranging from –8 (very poorly done), through 0 (no opinion), to 8 (very well done). Participants answered this question twice: once after reading the prolegalization study and once after reading the antilegalization study.

Information bias. We used the degree to which participants rated the article that supported their initial position as being of higher quality than the article contrary to their initial position as our measure of information bias (Klaczynski, Gordon, & Fauth, 1997; McHoskey, 1995). When participants rated the study congruent, as opposed to incongruent, with their initial attitude as higher in quality, they received positive information bias scores. When participants rated the study incongruent, as opposed to congruent, with their initial attitude as higher in quality, they received a negative information bias score.

Attitude polarization. To assess attitude change, four times during the exercise (after the brief statement of each study and after the detailed description of each study) we told
students the following: “Based on the preceding statement, rate how your opinion has changed toward legalizing marijuana.” Participants made their rating on a 17-point scale ranging from –8 (more opposed to legalization), through 0 (no change), to 8 (more in favor of legalization). By averaging over the four measurement occasions, we formed attitude-change scores.

We computed attitude polarization scores by multiplying attitude-change scores by 1 for prolegalization participants and by –1 for antilegalization participants. Participants whose attitudes changed in the direction of their initial attitude, indicating a more extreme attitude toward legalization of marijuana, received a positive-attitude polarization score. Negative scores indicated a less-extreme attitude after reading the studies.

Methodological reasoning. To assess methodological reasoning, we asked the participants to respond in writing to the following open-ended question: “Why do you think the (detailed) description of the study does or does not support the argument that marijuana has detrimental effects on society?” Two blind judges rated the critiques of the pro- and anti-articles on a 4-point scale ranging from 0 to 3. Zero responses showed no methodological reasoning, as revealed by unquestioning acceptance of the study’s findings. Responses that scored a 3 stated the need for control of variables to establish causality. The intraclass coefficient of scorer reliability was .75, which is acceptable for interrater reliability (Fleiss, 1986).

Findings

Methodological Reasoning

First, we compared methodological reasoning for students in the research methods course (trained group) with that of the comparison group. We combined the students in the introductory class and the tutors (untrained group) after a priori contrasts on information bias and attitude polarization revealed nonsignificant differences. Obtained scores for methodological reasoning ranged from 0 to 2, with the mean significantly higher for the trained group (M = 1.22, SD = 0.54) than for the untrained group (M = 0.55, SD = 0.41), t(175) = 9.11, p < .01. The effect size, indicating the standardized mean difference between the trained and untrained groups, was 1.46.

Study Quality

Next, we assessed the students’ evaluation of the quality of the information, irrespective of initial attitude and type of article. Trained students devalued both of the studies (M = –3.88, SD = 3.06). The untrained students were much less critical of the same studies (M = 0.1, SD = 1.47), t(175) = 8.93, p < .01, and showed a distinct preference for the article with which they agreed (see next section).

Information Bias

We tested for bias in information judgments on an absolute scale (i.e., relative to zero). For the untrained group, the mean for bias (M = 0.92, SD = 2.36) was significantly different from 0, t(117) = 3.29, p < .01. This result indicates a significant preference for the study in which the findings were in agreement with their initial attitude. In contrast, the trained group (M = 0.14, SD = 1.98) did not display significant bias, t(58) = 0.54, p > .05. Furthermore, we conducted a t test to examine differences in bias between the trained and untrained groups. We observed significantly less bias in information judgments, t(175) = –2.19, p < .05, in the trained group than in the untrained group. The effect size was .35.

Attitude Polarization

Measurements of attitude polarization closely paralleled our findings for information bias. The untrained group (M = 1.47, SD = 1.96) exhibited significant attitude polarization (i.e., relative to zero), t(117) = 5.75, p < .001. The trained group (M = 0.57, SD = 1.52) also reported significant attitude polarization, t(58) = 2.88, p < .01. However, a between-group comparison revealed that attitude polarization was significantly less in the trained than the untrained group, t(175) = –3.09, p < .01. The effect size was .49.

Mediation Analysis

To investigate whether methodological reasoning influenced informational bias, we conducted a mediational analysis of the relations among training, methodological reasoning, and information bias (Baron & Kenny, 1986). After coding the training variable (0 = untrained, 1 = trained), we found that methodological reasoning was significantly related to information bias, even after controlling for training, b(174) = –0.24, p < .01. Also, we observed that the effect of training on information bias was nonsignificant when methodological reasoning was controlled, b(174) = –0.09, p > .05. A z test (MacKinnon & Dwyer, 1993) on the correlation matrix indicated that the indirect effect of training on information bias, through methodological reasoning, was significant, z = –2.02, p < .05. In other words, methodological reasoning fully mediated the effect of training on biased information judgments.

Discussion

On the assumption that students versed in methodological reasoning would recognize obvious flaws in both studies, we anticipated that students in the research methods course would depreciate the information from both studies. We expected, then, they would not have any reason to change their initial attitude for or against legalization of marijuana. The finding of a modest amount of attitude polarization for the trained group was inconsistent with our expectations. However, the polarization in the trained group was significantly less than that in the untrained group, which was consistent with our expectations. Turning to students untrained in methodological reasoning, we expected they would accept the information at face value and exhibit biased assessment of information (i.e., judge studies congruent with their position to be better than studies in-
congruent with their position). Additionally, we anticipated the untrained group would use an “inflated” estimate of the evidence as a reason to change their attitude and display appreciable attitude polarization. Again, the results confirmed our expectations.

Psychologists have defined the fundamental cognitive bias as the propensity to overweight favorable evidence and underweigh opposing evidence (Fiske & Taylor, 1991; Nisbett & Ross, 1980). In this exercise, students untrained in practical applications of methodological reasoning exhibited this cognitive bias. The trained group, on the other hand, evinced this bias to a lesser degree. Instead, applying principles of methodological reasoning, they displayed a willingness to depreciate the flawed studies and testimony of authorities, regardless of the study’s conclusions. They also were less willing to rely on their own unsupported beliefs as a basis for changing their attitude.

These results are consistent with the notion that direct instruction can enable college students to resist biased information proffered by written communication. We point out that due to the uncontrolled threat to validity of previous exposure to psychology courses, the findings of the capstone exercise do not provide causal evidence of the effectiveness of the intervention. However, the positive findings obtained in the statistical mediational analysis do permit one to conclude with some confidence that skillful methodological reasoning does lessen the likelihood of biased information judgments. We recommend that investigators direct future research to these questions. If possible, they should include randomly assigned experimental and control groups. Alternatively, a pre–post nonequivalent control group design might prove useful.

Conclusions

Because psychologists know a great deal about how to teach the principles of scientific and statistical methodology (Lehman et al., 1988; Lehman & Nisbett, 1990; Mill, Gray, & Mandel, 1994), they are uniquely positioned to contribute to the reform of science education (National Research Council, 2000). Furthermore, the classroom experiences described in this article and in other work suggest that students who have the opportunity to construct useful scientific knowledge in an active-learning environment improve significantly in their ability to make informed decisions (Halpern & Nummedal, 1995). We hope this article will be useful to psychology instructors in their efforts to facilitate the development of analytical skills that students can apply in their daily lives.

References


Notes

1. This research was supported by a grant from the College of Liberal Arts and Sciences, Arizona State University.
2. We thank Jonathan M. Myers and Jeanine Polcastro for their help with data collection and analyses.
3. Send correspondence and requests for a complete set of materials to Barry Leshowitz, Department of Psychology, Arizona State University, Tempe, AZ 85287; e-mail: leshowitz@asu.edu.
Active and Passive Touch: A Research Methodology Project
Cynthia D. O'Dell and Mark Sudlow Hoyert
Teaching of Psychology 2002 29: 292
DOI: 10.1207/S15328023TOP2904_07
The online version of this article can be found at:
http://top.sagepub.com/content/29/4/292

Published by:
SAGE
http://www.sagepublications.com
On behalf of:
Society for the Teaching of Psychology
Additional services and information for Teaching of Psychology can be found at:
Email Alerts: http://top.sagepub.com/cgi/alerts
Subscriptions: http://top.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Oct 1, 2002
What is This?
Active and Passive Touch: A Research Methodology Project

Cynthia D. O’Dell
Departments of Women’s Studies and Psychology
Indiana University Northwest

Mark Sudlow Hoyert
Indiana University Northwest

We describe a perceptual experiment that we have successfully used in Research Methods classes. Students attempt to identify a series of simple cookie cutter shapes using only the fingers and hands (haptic perception). Students read archival studies that have used this procedure, identify confounds, generate and test alternative hypotheses, and present the results. Students described the project as quite enjoyable. They also showed a deeper understanding of research design and were able to present their data clearly. This exercise is an effective project for exploring and practicing critical analysis of existing research, experimental design, and data analysis, as well as data and theory presentation.

People can accurately identify objects using only the information provided by their fingers and hands, a process known as haptic perception. Gibson (1962) argued that form perception becomes increasingly accurate as people move their fingers and hands over an object. He identified two varieties of haptic perception, active and passive touch. Active touch roughly corresponds to the pattern of activity that people colloquially call touching. That is, active touch occurs when people move their fingers and hands to explore properties of the object. In contrast, passive touch does not involve movement of the hands and fingers. The stimuli are simply impressed on the skin. Our classroom activity explores differences between active and passive touch. This project joins a set of research methods exercises that replicate classic studies (e.g., Suter & Frank, 1986; Ware & Johnson, 1996). We find this project to be appealing, because the basic findings are easy to produce in a classroom setting, data collection proceeds quickly, and the existing literature is small and highly accessible. These advantages enable our students to expend more of their effort toward learning about research design and presentation. This project is our most successful exercise for introducing and practicing concepts and techniques in research design including (a) reading and understanding archival literature, (b) designing experiments with special emphasis on identifying and correcting confounds, (c) developing and testing hypotheses, and (d) presenting data and theory.

Gibson (1962) provided the first examination of differences between active and passive touch. He visually presented adult participants with a set of six cookie cutters and a drawing of those cutters. In all trials, the participants placed their dominant hand out of sight behind a screen and the experimenter placed a form in that hand. In the active condition, the participants cupped their fingers and felt the edges. In the passive condition, the participants kept their hand flat while the experimenter pressed the form into the palm. The participants were very accurate during the active touch condition with a mean frequency of 95% correct. This percentage dropped dramatically to 49% in the passive touch condition. Although Gibson’s original study contained several methodological flaws, subsequent research designed to correct some of the experimental confounds (e.g., Heller, 1980, 1984; Heller & Myers, 1983) has confirmed the general finding that active touch produces more accurate object recognition than passive touch. The research project we describe begins as a replication of the Gibson (1962) study, but moves beyond it as students redesign the study to eliminate confounds.

Preparation

Prior to the project, we ask our students to read the Gibson (1962) article. At this time, we assemble several sets of small (2.5 – 3.75 cm) metal cookie cutters for use as stimuli. We use five cookie cutters per set. Each set can serve as stimuli for four to five students. The particular shape is not an extremely important variable. However, the cookie cutters within any set should be approximately the same size and should be perceptually distinct from each other. For example, one of our sets contains the familiar shapes: moon, star, spade, flower, and heart. We constructed an object-matching page containing visual representations of the shapes. We trace each cookie cutter onto a sheet of blank paper and fill in the outline. We use a screen to obscure the vision of the participant. We constructed a wooden platform with a felt curtain that students can pass their hands through. However, a screen could be as simple as a piece of opaque fabric that two students stretch between them. Finally, we construct a data-coding sheet for each participant. Our sheets contain blanks for 20 active trials and 20 passive trials (five cookie cutters presented four times each) with places for order of presentation as well as for recording the actual response of the participant. At the top of each sheet, we record demographic information such as age, gender, and hand used in the study.
Basic Procedure

We ask students to work in groups of four or five. One person serves as the experimenter and presents the stimuli. We do not let this person serve as a participant because the experimenter manipulates the shapes extensively during presentation. Before the study begins, the experimenters fill out a coding sheet that specifies the order of stimulus presentation for each participant. Each experimenter generates a stimulus sequence in a semirandom order, with the limitation that a single stimulus does not appear more than twice in a row and that all stimuli are presented exactly four times. As a result, each participant experiences a different order of stimulus presentation. Furthermore, we ask the experimenters to group all 20 active and 20 passive trials together. Experimenters counterbalance the sequencing of active and passive trials across participants. The experimenters place the screen between themselves and the participants so that the participants cannot see their own hand or the stimuli. The participants place their hands through the screen with the palm up. The experimenters explain that the task is to try to identify the objects and then place the object-matching sheets in front of the participants. During active trials, the experimenters explain that the participants may move their palms and fingers around the objects. The experimenters place the objects in the participants' hands and remove them after 2 sec. During passive trials, the experimenters explain that the participants should hold their palms flat. The experimenters then press the objects gently into the palms for 2 sec. In all trials, the experimenters ask the participants to identify (from the object-matching sheet) the shape and then record their response. The data collection continues in this fashion until all trials are complete. Setting up and running the basic experiment takes about 15 to 20 min.

Replications

The basic experiment described is a fairly close replication of Gibson's (1962) original study. We have the students begin this project by collecting data using this procedure. The students compute the mean number correct and the standard deviation for the active and passive touch conditions. They also make a tally of the number of incorrect responses per stimulus to determine whether there is a pattern of mistakes.

After completing the first iteration of the study, the student groups discuss the study results, compare their results to Gibson's (1962) data, and identify any confounds they believe may have affected the data. The whole class then discusses the results and possible confounds and picks one of the generated suggestions for redoing the experiment. For example, our classes have proposed that the study should be repeated (a) using only the palms in both the active or passive touch conditions, (b) using only the fingers in both active and passive touch conditions, (c) making passive touch have movement without volition, (d) altering the time allotted for each trial, and (e) changing stimuli due to error patterns. These discussions provide an opportunity to introduce and explore many aspects of experimental design such as the importance of experimental control, the need to change only one independent variable at a time, order effects, and counterbalancing. Data collection proceeds quickly, as the testing of any one participant takes about 5 min. We devote two 75-min class meetings to this project and can run four or five permutations of the procedure in that time period.

Following each permutation, our students compute means and standard deviations. Following the final permutation, they compute means, standard deviations, and inferential statistics. We also then ask them to read a larger sample of the archival studies. The basic literature for the experiment includes the original Gibson (1962) study as well as a series of later studies that altered various aspects of the original procedure (Cronin, 1977; Heller, 1980, 1984; Heller & Myers, 1983; Schwartz, Perey & Azulay, 1975). These articles are all short and accessible, which greatly helps students understand the basic questions and procedures. The improved comprehension later helps students write clear and structured introduction and discussion sections for their American Psychological Association (APA) style experimental reports.

Across different semesters, we have asked students to read the literature both before and after data collection. We find that reading the Gibson article before the study and the rest of the articles after they have worked through several permutations of the procedure allows them to discover the confounds on their own. This process demonstrates to our students that they are capable of producing insightful ideas much like those of published scientists.

Assessment

Of all the experiments that our research methods students conduct, this one has proved to be their favorite. They give the projects a high score on the general course evaluation (M = 4.5, SD = 0.75) based on a scale ranging from 5 (outstanding) to 1 (poor). We also have the students evaluate the projects separately at the end of the semester and this project always receives the highest average ratings (M = 4.8, SD = 0.47) with the other project scores averaging ranging from 3.5 to 4.5, F(3, 147) = 18.32, p < .001.

This project also led to better comprehension of the course material. We require the students to write APA-style laboratory reports for each of the projects in the class. The papers describing this experiment are noticeably clearer than the others. We asked three faculty members (the course instructor and two others) to read and evaluate the final laboratory paper written by 38 students. Eighteen papers described the active and passive touch experiment. Twenty papers described two other studies (the effect of imagery on memory and the effect of goal orientation on academic success). The evaluators independently rated the papers on eight different criteria: (a) overall quality, (b) use of APA style, (c) organization of information, (d) comprehension and presentation of archival literature and issues, (e) comprehension and presentation of methods, (f) comprehension and presentation of results, (g) integration of current data into a theoretical framework, and (h) discussion of future directions. They assigned a score for each criterion that ranged from 0 to 100 where higher scores
indicated better performance. The faculty evaluations were quite consistent, \( r(326) = .86 \), for Readers 1 and 2; \( r(62) = .83 \), for Readers 1 and 3; and \( r(62) = .82 \), for Readers 2 and 3. We compared the scores using a MANOVA. As shown in Table 1, the overall quality of the two sets of papers was comparable, as was the use of APA style, the comprehension and presentation of results, integration of current data into a theoretical framework, and discussion of future directions. However, the evaluations for organization of information, comprehension and presentation of archival literature and issues, and comprehension and presentation of methods were significantly higher for the papers describing the active and passive touch study than for the other experiments. We suspect that the advantage is accrued because the literature and the theoretical issues are more accessible, thus enabling our students to write about a topic that they more fully comprehend. Working through confounds step-by-step also aids the understanding of methodology. We can see this effect by looking at course quiz performance. Students who had completed the active and passive touch experiment scored better on the quizzes testing methodological issues addressed within this project than students who had explored these same issues using other experiments (\( M = 81.0, SD = 10.6 \) vs. \( M = 74.4, SD = 17.3 \)), \( t(48) = 2.57, p = .013 \).

## Conclusions

We have used this experiment for several years in our Research Methods courses and have found the project to be quite valuable in introducing and practicing concepts and techniques in research design. We have also adapted this project for use in a Sensation and Perception course. When we do so, we integrate the project into the larger framework of Gibson’s (1966) perception and action theory. We discuss possible neurophysiological evidence distinguishing active and passive touch (Sakata & Iwamura, 1978). In both Sensation and Perception and Research Methods courses, we have found that the project is successful in engaging students in developing and applying skills and knowledge. As such, it embodies the best characteristics of an active learning technique (Mathie et al., 1993; Ware & Johnson, 1996).

## References


## Note

Send correspondence to Cynthia O’Dell, Indiana University Northwest, Departments of Women’s Studies and Psychology, 3400 Broadway, Gary, IN 46408; e-mail: codell@iun.edu.

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Touch M</th>
<th>Touch SD</th>
<th>Other M</th>
<th>Other SD</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall quality</td>
<td>83.4</td>
<td>7.6</td>
<td>80.8</td>
<td>9.9</td>
<td>0.65</td>
<td>.426</td>
</tr>
<tr>
<td>Use of APA style</td>
<td>87.8</td>
<td>7.4</td>
<td>83.9</td>
<td>10.5</td>
<td>1.91</td>
<td>.176</td>
</tr>
<tr>
<td>Organization</td>
<td>91.3</td>
<td>4.7</td>
<td>83.7</td>
<td>11.9</td>
<td>6.79</td>
<td>.014</td>
</tr>
<tr>
<td>Archival literature</td>
<td>89.4</td>
<td>6.5</td>
<td>78.6</td>
<td>15.5</td>
<td>6.75</td>
<td>.014</td>
</tr>
<tr>
<td>Methods</td>
<td>80.9</td>
<td>4.9</td>
<td>80.8</td>
<td>9.9</td>
<td>15.53</td>
<td>.001</td>
</tr>
<tr>
<td>Results</td>
<td>80.8</td>
<td>10.0</td>
<td>77.9</td>
<td>9.6</td>
<td>3.09</td>
<td>.088</td>
</tr>
<tr>
<td>Theoretical framework</td>
<td>80.0</td>
<td>10.4</td>
<td>77.4</td>
<td>11.1</td>
<td>1.99</td>
<td>.168</td>
</tr>
<tr>
<td>Future directions</td>
<td>75.7</td>
<td>11.8</td>
<td>71.3</td>
<td>16.1</td>
<td>0.90</td>
<td>.349</td>
</tr>
</tbody>
</table>

Note. APA = American Psychological Association.
References for Core Texts by Category With Series and Price Information

Review Texts

Extended Review Texts

Outline Texts

* These two 2-volume core texts are partially identical. The super review volumes are the same as the essentials volumes but also include some of the problem/solution discussions and practice multiple-choice self-test questions from REA’s Psychology: A Complete Solution Guide to Any Textbook (Staff of REA, 1999). We included both 2-volume texts because some teachers might prefer the super review books because of their additional material.

Notes
1. We thank Randolph Smith and three anonymous reviewers for their comments on an earlier version of this article.
2. Send correspondence to Richard A. Griggs, Department of Psychology, PO Box 112250, University of Florida, Gainesville, FL 32611; e-mail: rgriggs@ufl.edu.

Using the Barnum Effect to Teach Psychological Research Methods

Thomas E. Boyce
University of Nevada, Reno

E. Scott Geller
Virginia Polytechnic Institute and State University

We used the Barnum Effect (Dickson & Kelly, 1985) to teach principles of research and to promote participation in 2 undergraduate psychological research methods courses. During the first week of class, 2 instructors presented themselves as experts in graphology and collected handwriting samples from students. Subsequently, each student received the same ambiguous one-size-fits-all personality assessment, presumably based on his or her handwriting. The students completed questionnaires about their perceptions of graphology prior to and immediately after receiving their personality assessments and 8 weeks later. Students significantly increased their ratings of graphology as a science after receiving their personality assessments, then lowered their ratings significantly after debriefing. We used the demonstration to teach various aspects of displaying and interpreting data and to highlight the pitfalls of pseudoscience.

The Barnum Effect refers to people’s tendencies to accept general and ambiguous statements as descriptive of their unique personalities (Dickson & Kelly, 1985). It was named after P. T. Barnum, the famous circus entrepreneur, who claimed he owed his success to two basic principles: “There’s a sucker born every minute” and “have something for everybody” (Myers, 1993, p. 153). Both of these principles seem to be verified by people’s acceptance of one-size-fits-all personality statements.

Dickson and Kelly (1985), for example, administered a personality test to participants, provided them the same bogus personality profile presumably generated from the personality test they took, and then had participants rate the perceived accuracy of the profile they received. Most participants rated the general personality profile as a good or excellent description of themselves. Furthermore, participants’ subjective ratings increased directly with the favorableness of the general profile.

The literature is replete with examples of applications of the Barnum Effect to teach ethics in research (Beins, 1993) and to identify variables that moderate its effect (for a review, see Dickson & Kelly, 1985). However, we found no published research on the use of the Barnum Effect to promote healthy skepticism of pseudoscience and teach university students research design and methodology. Here we document the results of a Barnum procedure designed to accomplish these objectives.

We selected graphology as the personality assessment procedure because of people’s tendencies to assign higher accuracy ratings to profiles generated by more ambiguous assessment procedures (Dickson & Kelly, 1985). Our aims were to demonstrate the allure of and reduce students’ belief in pseudoscience as well as develop an appreciation for em-
pirical research over intuition or common sense. The latter goal is consistent with the work of Miller, Wozniak, Rust, Miller, and Slezak (1996), who demonstrated that writing an essay inconsistent with beliefs was more effective at reducing students’ erroneous beliefs about psychology than either reading about a topic or reading an essay that stated an appropriate argument that was contrary to their beliefs.

Method

Participants and Setting

Participants were 74 undergraduate college students enrolled in two sections of a Principles of Psychological Research course at a large state university. One class, taught by the senior author, enrolled nonpsychology majors, whereas the other, taught by the second author, was for psychology majors only. The students ranged in age from 18 to 45 years with a mean of 20. Women outnumbered men about two to one. The course attracted primarily college sophomores and was a required course for psychology majors. Twenty-six students were psychology majors, and 48 were nonpsychology majors.

Materials

A written questionnaire containing 10 Likert-type questions assessed students’ perceptions of graphology as a science. Students rated each question on a 7-point Likert-type scale with anchors at 1 (strongly disagree) and 7 (strongly agree). Two items were reverse scored. In addition, the written questionnaire asked students to rate, on a scale ranging from 1 (definitely NOT scientific) to 10 (definitely scientific), how scientific 10 disciplines or knowledge domains were, including physics and psychology.

A one-size-fits-all personality profile, adapted from Forer (1949), provided the illusion that personalities could be identified from simple handwriting samples. The profile appears in the Appendix.

Procedures

We administered questionnaires three times during regular class meetings.

Pretest. We administered the first questionnaire during the second meeting of each class. Afterward, we presented ourselves as experts in the science of graphology. We discussed our interest and expertise in graphology and obtained handwriting samples by asking students to write on a blank sheet of paper “The quick brown fox jumps over the lazy dog.” The instructions explained that this sentence contained every letter in the English alphabet and therefore enabled a comprehensive writing sample. The students wrote their normal signature three times anywhere on the paper and then wrote a fourth signature on the reverse side of the paper as slowly as possible.

Posttest. Prior to administering the posttest questionnaire on the third meeting of each class, we gave the same bogus one-size-fits-all personality profile individually to each student, along with their handwriting sample. To provide the illusion of an analysis, the handwriting sample included various irrelevant letters and symbols in red ink. We customized the profiles for each student with his or her name.

Immediately after allowing students to read their personality profile, we administered the questionnaire described previously. Then, we discussed the personality profiles openly in class and debriefed the students regarding the true purpose of the exercise.

Follow-up. We administered the questionnaire in class once again during the last week of class (8 weeks after the posttest administration).

Results

Mean ratings of graphology as a science (measured on a 7-point scale) for each experimental phase appear in Table 1. We tracked repeated measures by the last 4 digits of students’ social security numbers followed by their middle initials. A 2 (class: psychology majors vs. nonmajors) × 3 (phase: pretest, posttest, follow-up) ANOVA tested for differences in students’ perceptions of graphology as a science. A main effect for phase occurred, $F(2, 181) = 122.21, p < .0001$. Neither the main effect for class nor the Class × Phase interaction was significant, $p > .40$, respectively. Thus, we collapsed data across classes for post hoc analyses.

Planned pairwise comparisons of the graphology ratings using Fisher’s least significant difference revealed that the pretest mean of 4.3 was significantly lower than the posttest mean of 5.0 and was significantly higher than the follow-up mean of 2.7. The difference between the posttest and follow-up ratings was also significant, all $p < .0001$. As expected, perceptions of physics and psychology (used as controls) did not vary significantly as a result of the graphology demonstration or other classroom activities. These results reflect that, in both classes, students’ perceptions of graphology, a pseudoscience, increased after receiving the personality profile and declined after being debriefed. In contrast, their perceptions of legitimate sciences did not change significantly.

<table>
<thead>
<tr>
<th>Group</th>
<th>Pretest</th>
<th>Posttest</th>
<th>Follow-Up</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majors only</td>
<td>4.2</td>
<td>5.1</td>
<td>2.8</td>
</tr>
<tr>
<td>Nonmajors</td>
<td>4.4</td>
<td>4.9</td>
<td>2.5</td>
</tr>
<tr>
<td>Overall</td>
<td>4.3</td>
<td>5.0</td>
<td>2.7</td>
</tr>
</tbody>
</table>

Note. Judgments were based on a 7-point scale ranging from 1 (strongly disagree) to 7 (strongly agree).
Discussion

Consistent with previous findings on the effectiveness of counterattitudinal advocacy to reduce students’ erroneous beliefs (e.g., Miller et al., 1996), the Barnum demonstration used here successfully reduced students’ beliefs that graphology is a valid science. In addition, we used the graphology results to teach interpretation of data and documentation of research for professional presentation. The absence of any significant differences due to instructor or class suggests the effects of the teaching methods generalized across psychology majors and nonmajors.

It was important that the graphology demonstration occur during the first week of class. At this time, the students had little history with the instructors and the class emphasis on acquisition of knowledge via the scientific method. The introductory chapter of the textbook (Goodwin, 1998, pp. 5–6) discussed “faulty ways of knowing,” one of which was “authority,” an effect that could underlie the graphology demonstration (Hinrichsen & Bradley, 1974). Thus, debriefing students appropriately at the onset of the semester established a learning context for students to challenge their common sense, become wary of pseudoscience, and think critically about daily information they receive. In other words, like the demonstration by Miller et al. (1996), we used this exercise to set the stage for mindful learning (Langer, 1997), which we attempted to sustain throughout the semester.

Specifically, during the debriefing we assured the students the graphology exercise was the only classroom demonstration that would involve deception. We clearly stated that the remaining information conveyed would be state-of-the-art of psychology as an empirical science. We explained that during the remainder of the semester we would discuss the research process by which scientists come to understand environment–behavior relations and develop psychological theory and principles.

However, to facilitate an understanding of the research process, we did refer to the graphology findings throughout the semester to teach students about ethics in research, including the use of deception and the importance of debriefing (e.g., Beins, 1993). In addition, we displayed the data graphically to demonstrate the appropriate use of line versus bar charts and explain how to determine main effects and interactions from a pictorial representation of group means. These activities appeared to engage students because it was their scores making up the data being displayed.

As members of teaching/learning teams, the students conducted research activities of their own. They also prepared written documentation of their team effort and orally reported on their research projects on the last day of each class. During the preceding class, we formally presented the graphology study to (a) provide students with a model by which to prepare their own oral presentations, (b) reemphasize the importance of critically evaluating research-relevant information, (c) demonstrate key principles taught throughout the semester, and (d) ensure that no students remained concerned over having been deceived.

It is noteworthy that five students reported their intention to switch their majors to psychology during the next school semester as a result of their participation in the class. Thus, it appears the classroom demonstration described here was pedagogically useful for both the students and the instructors, who received overall mean course ratings of 3.7 and 3.9 on a 4-point scale with 4 anchored at outstanding. We hope our successful use of the Barnum demonstration will assist other instructors in challenging students’ common sense with creative classroom activities.

References


Appendix

Personality Profile Presented to All Students
After Obtaining Pretest Questionnaire Data and Writing Samples

You have a strong need for other people to like and to admire you. You have a tendency to be critical of yourself. You pride yourself on being an independent thinker and do not accept other opinions without satisfactory proof. You have found it unwise to be too frank in revealing yourself to others. At times you are extraverted, mild-mannered, and social; at other times you are introverted, wary, and reserved. Some of your aspirations tend to be pretty unrealistic.

It seems you are nursing a grudge against someone; you really ought to let that go. You worry about things more than you let on, even to your best friends. You are adaptable to social situations and your interests are wide ranging.

Note

Send correspondence and requests for a copy of the questionnaire used in the study to Thomas E. Boyce, Department of Psychology/296, University of Nevada, Reno, NV 89557; e-mail: teboyce@unr.edu.
Science or Snake Oil? Teaching Critical Evaluation of "Research" Reports on the Internet
Patricia A. Connor-Greene and Dan J. Greene
Teaching of Psychology 2002 29: 321
DOI: 10.1207/S15328023TOP2904_14

The online version of this article can be found at:
http://top.sagepub.com/content/29/4/321
Science or Snake Oil? Teaching Critical Evaluation of “Research” Reports on the Internet

Patricia A. Connor-Greene
Clemson University

Dan J. Greene
Department of Mental Health
Anderson, South Carolina

The proliferation of information on the Internet introduces new challenges for educators. Although the Internet can provide quick and easy access to a wealth of information, it has virtually no quality control. Consequently, the Internet has rendered faculty more essential than ever as teachers of the analytic and evaluative skills students need to become educated consumers of information. In this article we describe an exercise using small-group discussion and individual problem-based learning to teach critical thinking about the Internet. Data from the exercise and from student evaluations support both its need and students’ perceptions of its effectiveness.

The abundance of information available on the Internet presents both benefits and challenges to its users. Although the Internet provides immediate access to a wealth of data, it has no gatekeeper (Munger, Anderson, Benjamin, Busiel, & Paredes-Holt, 1999). In contrast to journal articles that are peer reviewed to ensure credibility and detect inaccuracies (Brandt, 1996), anyone can post anything to the Internet, whether the information is based on solid evidence or total fabrication (Peters & Sikorski, 1997). The lack of quality control makes the Internet “the world’s largest vanity press” (Silberg, Lundberg, & Musacchio, 1997, p. 1244) and the host of a great deal of misinformation (Kassirer, 1995; Regan, 1998; Smock, 1995). The Internet has also become an extremely popular marketing tool. As Silberg et al. (1997) commented, “At first glance, science and snake oil may not always look that different on the Net” (p. 1244).

One of the harmful aspects of the Internet is its unchecked distribution of health information. Unlike most health care professions that require a license to practice, the Internet fosters “indiscriminate credentialism” in which someone with no formal training can present oneself as an expert while providing erroneous information (Bader & Braude, 1998, p. 409).

Critique and evaluation are essential components of becoming a savvy consumer of research (Halpern, 1998; Lawson, 1999; Sternberg, 1999). Some excellent resources are available to help the public evaluate information on the Internet (e.g., http://www.devry-phx.edu/lnresrc/dowsc/integrity.htm). To assist consumers in finding credible health care information, the federal government has created a site (http://www.healthfinder.gov/) that lists government agencies, non-profit organizations, and self-help Internet sites. HealthWeb (http://healthweb.org/) is another directory of legitimate health-related Internet sites (Bader & Braude, 1998).

The overwhelming volume of information available on the Internet has rendered faculty more essential than ever as teachers of the analytic and evaluative skills students need to become critical consumers of enormous quantities of information. A “teachable moment” presented itself when several students received a mass e-mailing about the dangers of aspartame. The e-mail referenced a Web site (http://www.rense.com/health/sweetners.htm). We used an article from this Web site in a class exercise on thinking critically about the Internet.

Class Exercise

In two upper level undergraduate psychology courses, students read an article (Markle, 1998) printed from an Internet site that warned of the dangers of aspartame. After reading this article in class, students completed an anonymous questionnaire in which they wrote their reactions to the article. After students completed and turned in the questionnaires, the class broke into small groups of three to four students to discuss the following questions:

1. After reading this article, what advice, if any, would you give to a friend who drinks three diet sodas every day? Why?
2. Does the author make a convincing case that aspartame causes health problems? Why or why not?
3. Are there any “red flags” that lead you to question the credibility of this article? (Each group made a written list of any red flags they identified and handed this in at the end of the group exercise.)
4. What is the nature of evidence presented by the author?
5. Does the evidence warrant the conclusion that aspartame causes health problems?
6. After discussing this article, would it be appropriate to conclude that aspartame is dangerous? Would it be appropriate to conclude that it is safe?

7. Even if you knew nothing about biochemistry, how would you evaluate the credibility of the claims in this article? What sources of information would be useful to you in this process?

Class Discussion

After students discussed the questions in small groups (approximately 30 min), we reconvened for a class discussion. Students readily volunteered a variety of problems they noticed in the article that led them to question its credibility. These red flags included no credentials listed for both the author and the “experts” cited in the article, no research evidence cited to support allegations, no citations for sources, an extensive “laundry list” of symptoms allegedly caused by aspartame (e.g., headache), sensationalized writing style, and sweeping statements (“aspartame destroys the nervous system,” Markle, 1998, p. 4).

As both a red flag and in response to Question 4 concerning the nature of the evidence provided by the author, students identified a variety of weaknesses in the data. Virtually all the evidence cited in the Markle (1998) article was anecdotal, circumstantial, or hearsay, insufficient to allow verification (e.g., “During a visit to a hospice, a nurse said that six of her friends, who were heavy Diet Coke addicts, had all been diagnosed with MS. This is beyond coincidence”; Markle, 1998, p. 2). The article contained no reports of controlled research studies to support the alleged causal relation between aspartame and health problems. Furthermore, the author cited no statistical data to support even a correlational relation between aspartame and illness.

The class generated a list of possible sources to use to check the information in the Markle (1998) article that would be accessible to someone without extensive knowledge of biochemistry (Question 7). Their list included Web sites for organizations that specialize in the diseases mentioned in the article (e.g., multiple sclerosis, diabetes, Alzheimer’s), PsycLIT and Medline database searches on aspartame, and searches on Markle and any experts she cited by name.

After the class generated a list of possible sources of information on aspartame, I (PCG) provided an overview of the extensive research findings on aspartame and health, which did not support Markle’s (1998) extreme statements about aspartame as a deadly substance.

Analysis of Student Questionnaire Data

Of the 39 students who participated in this exercise, 23 (58.9%) reported using products containing aspartame. Fifteen of the students (38.5%) reported that they had previously heard or read about problems with aspartame; several identified multiple sources of this information. Among those students who had previously heard of aspartame dangers, 2 students (13%) had received a recent e-mail message about it, 14 (93%) had heard about it “from someone else,” 1 (6%) reported seeing it on television, 2 (13%) reported reading it in a newspaper, and 2 (13%) reported reading it in a magazine. None of the students who reported learning about it from television, a magazine, or a newspaper could recall the name of the specific source (television program, name of magazine or newspaper).

We independently read and classified students’ initial written reactions to the article using three categories: (a) unquestioning acceptance (expressing shock, alarm, or concern without any mention of doubt, e.g., “I’ll never drink another Diet Coke!”), (b) ambivalence about whether to believe the article (e.g., “It’s kind of shocking—I wonder whether it’s true or not”), or (c) disbelief (e.g., “just a lot of hype”). We used Cohen’s (1960) kappa coefficient to assess interrater agreement (κ = .74). We classified the reasons for disbelief using the following categories: (a) vague reasoning with no specific reasons articulated (e.g., “seems a little far-fetched”; “seems very sensationalized”), (b) personal anecdotal evidence (e.g., “Most of my friends use Nutrasweet and none of them have MS”), (c) reliance on external authority (e.g., “if this was true, the FDA [Food and Drug Administration] would have taken aspartame off the market”), and (d) specific shortcomings noted (e.g., lack of causal evidence; κ = .82). We resolved any discrepancies in classification through discussion.

Twenty-one of the students (53.8%) unquestioningly accepted the information in the article, 10 (25.6%) expressed ambivalence, and 8 (20.5%) did not believe the article. Among the 8 students who did not believe the article, 2 used vague reasoning, 1 used personal anecdotal evidence, and 1 relied on external authority.

Only 4 of the 21 students (50% of those who disbelieved the article; 10.3% of the entire group of students) cited specific shortcomings in the article, including lack of causal evidence (1 student); argument based on emotion and “scare tactics” rather than data (1 student); no credentials given (1 student); sweeping, extreme statements (1 student); long laundry list of common symptoms (4 students); lack of statistical data (1 student); lack of reference information (2 students); and location on a commercial Web site (1 student).

Data From Small-Group Discussions

In contrast to individual responses in which only four students (10.3% of the class) cited specific problems with the article, in the small-group discussion phase of the exercise, all 11 groups of three to four students generated lists of specific red flags (M = 7.27, SD = 2.01; range = 5 to 11).

I (PCG) used the Markle (1998) article and structured small-group discussions in two consecutive semesters. In the second semester, rather than giving students an overview of aspartame research at the end of class, I gave an out-of-class individual assignment. I asked students to imagine that they were employed by the FDA; their assignment was to write a one to two page position paper on whether to take aspartame off the market. This assignment counted as a 10-point essay question on their first test. I based their grade on the credibil-
ity of their sources and the extent to which the research evidence they found warranted their conclusions.

After students brought their papers to class, they discussed their sources and conclusions in small groups prior to a larger class discussion. Overall, the students did an excellent job with the assignment; most papers reflected thorough, thoughtful searches of credible data sources and a good grasp of information. In class discussion, many students reported that they had searched sources they had never previously used (e.g., Medline, PsycLIT, FDA), and many talked about the challenge of sifting through large amounts of information to determine what is credible. In describing strategies they used to search and evaluate information, most students found it helpful to first obtain the “big picture” from a variety of credible sources, look for convergence in ideas from these sources to form the basis for their position, and then select specific representative articles to support their position. Seeking consistent findings from multiple-credible sources was a strategy that we had discussed in class prior to the position paper assignment.

Student Evaluations of the Exercise

Students completed evaluations of the exercise both semesters. Table 1 lists the evaluation results, contrasting the first semester (small-group discussion of Markle’s, 1998, article) with the second semester (small-group discussion of the article plus researching and writing individual position papers on whether aspartame is safe). Overall, the student ratings were quite positive. The classes that researched and wrote position papers in addition to group discussion rated the exercise significantly higher in helping them to think critically about the article and in helping prepare them for evaluating research data from other sources.

I honestly never really like outside work like this, but I actually really loved this project. I learned so much and I found myself getting more and more information just because I wanted to know more. I think I liked it so much because most projects don’t relate to us and this one really did. It was great!

In their initial responses to the Markle (1998) article, only one fifth of the class expressed disbelief, and of those students, half used vague or ad hominem reasoning or reliance on external authority to explain their doubt. Only 4 out of 39 students cited a specific reason for questioning the validity of the article, suggesting that even those who did have doubts had difficulty articulating their concerns. Only one student mentioned Markle’s lack of causal evidence, a key concept used in evaluating the strength of a research claim, and not a single student noted the anecdotal, circumstantial nature of Markle’s evidence. These data highlight the need for greater emphasis on teaching students to (a) question what they read and (b) develop the skills needed to clearly articulate the strengths and weaknesses of reports of research.

The students’ initial lack of critical thinking is disconcerting, especially for upper level psychology students. This article contained numerous red flags, yet the majority of students did not question its validity. An article with erroneous information but a stronger veneer of credibility would present even more challenges to its readers. Conversely, an article could contain some of the problems of the Markle (1998) article (e.g., spelling errors and emphatic language and punctuation) but still contain credible information. Surface cues are noteworthy, but to actually assess an article’s credibility, the evaluation must go deeper to see how well its claims fit with research data from other sources.

In contrast to the majority of the individual students’ responses indicating unquestioning acceptance of the article, every group noted and described multiple limitations. One

### Table 1. Student Evaluations of the Exercise

<table>
<thead>
<tr>
<th>Evaluation</th>
<th>Discussion Only&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Discussion/Position Paper&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>1. Helped me think more critically about this article</td>
<td>4.23</td>
<td>1.01</td>
</tr>
<tr>
<td>2. Will help me evaluate reports of research in the future</td>
<td>4.26</td>
<td>0.82</td>
</tr>
<tr>
<td>3. Helped me think about specific ways to evaluate credibility</td>
<td>4.44</td>
<td>0.60</td>
</tr>
<tr>
<td>4. Increased my awareness of potential problems of the Internet</td>
<td>4.31</td>
<td>0.83</td>
</tr>
<tr>
<td>5. This exercise was interesting</td>
<td>4.23</td>
<td>0.91</td>
</tr>
<tr>
<td>6. It was helpful to work in groups for this exercise</td>
<td>4.05</td>
<td>1.15</td>
</tr>
</tbody>
</table>

<sup>a</sup>N = 39, <sup>b</sup>N = 45. <sup>*p < .05</sup>, <sup>**p < .005</sup>.

### Table 2. Written Comments From Students About the Exercise

I'm really a guinea pig for these things. I totally believed this article at first. I'm going to think more about things I read. This really opened my eyes about what is published on the Internet.

People (myself included) are way too gullible. I have learned that credible sources and scientific evidence are necessary to make an argument that should be believed.

I learned new criteria for judging sources, claims, and scientific studies and became more familiar with the type of critical thinking that ought to be done when analyzing a claim.

I honestly never really like outside work like this, but I actually really loved this project. I learned so much and I found myself getting more and more information just because I wanted to know more. I think I liked it so much because most projects don’t relate to us and this one really did. It was great!
of the questions posed for the small-group discussion asked about red flags, alerting groups to look for problems when they might not have otherwise done so. Students received no such prompt for the individual response part of the exercise, which may explain the difference between levels of critical thinking in individual and group responses. Perhaps if students know there is a problem they can identify what it is, but they may not be initially alert to the possibility of flaws in what they read. Teachers must find effective ways to heighten students’ awareness of the need to think critically whenever they read, not just in response to specific prompts asking them to identify weaknesses.

The Internet provides an ideal forum for teaching students to evaluate and select useful, credible sources from a vast array of unfiltered information. Because students rely so often on the Internet, it is imperative that they become educated consumers of this medium.

This exercise evolved over two semesters to include a problem-based learning component in which students located and evaluated research themselves rather than being given this information by the teacher. Results of student evaluations of the exercise over both semesters suggest that this added component significantly strengthened students’ perceptions of the value of the exercise in teaching critical thinking. Part of our identity as good teachers may involve giving “gifts” of information, but in doing so, we may sometimes miss an opportunity for students to develop new skills by finding the information themselves. One of the challenges to teachers in fostering critical thinking and problem solving is changing our role from information-giving expert to facilitator. When this happens, a shift occurs; the emphasis is on what they learned rather than what we taught them.

Students’ written comments suggest the exercise was effective in helping them recognize their need for careful, critical reading, questioning, and verification. The topic’s relevance to their lives increased their investment in researching it, and the position paper required students to evaluate rather than merely summarize information. Students learned specific ways to evaluate credibility and located and used sources that were new to them.

Overall, the data from this exercise support both the need for and the perceived value of assignments that encourage careful, thorough reading and evaluation. Critical thinking is not an academic fad; it is an essential skill for living in the information age.

References


Note

Send correspondence to Patricia A. Connor-Greene, Department of Psychology, Brackett 410G, Clemson University, Clemson, SC 29634–1355; e-mail: connorg@clemson.edu.
Eight students responded to the open-ended question. One student said, “I like the system,” and another student noted, “I very much enjoyed having instant feedback and would like to see the current self-grade system continue.” In addition, another student noted the pedagogical advantage of immediate feedback, saying, “I liked actually doing my own grades b/c it made me more aware of what I got wrong.” Regarding the accuracy of the self-scoring, one student said, “I think having us copy our answers onto a separate sheet makes us feel like we can’t cheat—or if we do, we’ll be more likely to get caught.”

Discussion

These findings may be of interest to instructors who want to provide their students with the pedagogical advantage of immediate scoring on quizzes, while being confident that the scoring is accurate. Our data suggest that using a declaration sheet when students self-score their quizzes assures that students will accurately score their quizzes and accurately report their scores. Instructors may thus follow our practice of not checking each student’s actual quiz answers against his or her declaration sheet answers for each quiz. As such, this technique offers instructors a useful option for administering and scoring short multiple-choice or true–false quizzes.

Our data do not address the degree of accuracy in student self-scoring without the use of a declaration sheet. We do not have data from a “control” class that scored their quizzes without the declaration sheets or some other type of safeguard in place. Although it is not immediately evident how to determine the scoring accuracy of such quizzes, it may be that accuracy in self-scoring would be just as high if students did not use declaration sheets. Evidence suggests, however, that this outcome would be unlikely (Davis et al., 1992).

Another consideration in the use of the declaration sheet is the possibility that it may create a perception that the instructor does not trust the students. We always stress to students that we believe most students are honest, but that we also realize a small percentage of students may try to take advantage of self-scoring to unfairly enhance their score. Thus, we encourage students to think of the declaration sheet as a way for them to feel confident that other students are not doing anything inappropriate in the self-scoring of quizzes. Students typically accept this explanation, and some students have commented that they are glad the safeguard is in place.

This approach to quiz scoring provides an instructor the obvious advantage of reduced grading time. A recent longitudinal study on faculty workload and time allocation indicated that faculty now spend more time on both research and teaching than they did previously (Milem, Berger, & Dey, 2000). Given these increased time demands, allowing students to score their quizzes in class is one way that some faculty may choose to afford themselves more time for other responsibilities.

References


Notes

1. We thank Randolph Smith and three anonymous reviewers for helpful comments and suggestions on an earlier draft of this article.

2. Send correspondence to David M. Carkenord, Department of Psychology, Longwood University, Farmville, VA 23909; e-mail: carkenorddm@longwood.edu.

Testing Pseudoscientific Claims in Research Methods Courses

Frank M. LoSchiavo and Kristin L. Roberts
Ohio University–Zanesville

We describe a 10-week research methods course that introduced students to research design, data collection, statistical analysis, and manuscript preparation. Rather than focusing on lectures and brief activities, the course immersed students in a professional-grade, quarter-long laboratory experiment designed to empirically investigate the effectiveness of computer software that ostensibly detects deception through voice stress analysis. We discuss the relative merits of comprehensive laboratory projects, focusing specifically on investigations that test dubious, pseudoscientific claims. Furthermore, we highlight a basic schedule that enables instructors to incorporate full-scale laboratory projects successfully into their 10-week methods courses.

Research methods courses are ubiquitous (Messer, Griggs, & Jackson, 1999), yet difficult to teach. Consequently, Teaching of Psychology has published more articles on research methods than on any other topic (Johnson, 2001). Contributors have suggested a variety of course designs, most of which have focused on a mix of lectures, demonstrations, and brief research projects (e.g., Benedict & Butts, 1981; Brems, 1994; Carroll, 1986; Fontes & Piercy, 2000; Kerber, 1983; Sommer & Sommer, 2003; Underwood, 1975; Yoder, 1979). Although brief projects have several advantages, we
believe they risk trivializing psychological science and may implicitly teach students that they are incapable of conducting more rigorous, professional-grade research.

In contrast, several authors have described projects that required students to parallel the work of professional research psychologists. Whereas some projects have required students to replicate previous research (Chamberlain, 1988), others have required students to design new studies (Chamberlain, 1986; Pury, 2001), typically involving survey methods (e.g., Chapdelaine & Chapman, 1999; Froese, Vogts-Scribner, Ealey, & Fairchild, 2003; Marek, Christopher, & Walker, 2004). Although a few articles have featured realistic laboratory experiments, many have required the use of relatively expensive animal colonies for conditioning experiments (Michael, 1975) and surgical procedures (McGill, 1975).

This article focuses on a professional-grade, quarter-long laboratory experiment in which students tested dubious, pseudoscientific claims regarding Truster, a new computer program that ostensibly detects deception via voice stress analysis (Meyer, 1998; Taylor, 2002). Although the manufacturer reported that Truster had proven effective in controlled tests (Truster, 1997; Van Damme, 1998), several warning signs forced us to view their evidence with skepticism (Ruscio, 2002). For example, the manufacturer placed significant emphasis on anecdotal information, and their astonishingly positive empirical findings have never been replicated or published in a peer-reviewed journal. Furthermore, the manufacturer failed to address critical theoretical issues, such as whether a unique physiological lie response exists (Hollien, 1990; Lykken, 1998). Thus, it is not clear if the software can detect voice stress that is specific to telling lies.

Because most pseudoscientific and parapsychological topics are inherently interesting, they are particularly well suited for testing in methods courses (Lilienfeld, Lohr, & Motier, 2001). Students are likely to harbor pseudoscientific beliefs (Feder, 1985; Shermer, 1998), and instructors are eager to help them think more critically (e.g., Connor-Greene, 1993; Lawson, 1999, 2003). Indeed, Cole (1982) probably spoke for most instructors when he stated that the undergraduate curriculum should “seek to liberate the student from ignorance ... and from standing helpless before extravagant pseudoscientific assertions about human behavior” (p. 25).

Course Description

Precourse Preparation

Because review boards are slow to grant approval, the instructor (the first author) submitted a preliminary proposal that outlined several methods featured in the deception-detection literature (e.g., guilty knowledge test, mock theft). In addition, the instructor conducted a comprehensive literature search and created a local information archive that many have required the use of relatively expensive animal colonies for conditioning experiments (Michael, 1975) and surgical procedures (McGill, 1975).

This article focuses on a professional-grade, quarter-long laboratory experiment in which students tested dubious, pseudoscientific claims regarding Truster, a new computer program that ostensibly detects deception via voice stress analysis (Meyer, 1998; Taylor, 2002). Although the manufacturer reported that Truster had proven effective in controlled tests (Truster, 1997; Van Damme, 1998), several warning signs forced us to view their evidence with skepticism (Ruscio, 2002). For example, the manufacturer placed significant emphasis on anecdotal information, and their astonishingly positive empirical findings have never been replicated or published in a peer-reviewed journal. Furthermore, the manufacturer failed to address critical theoretical issues, such as whether a unique physiological lie response exists (Hollien, 1990; Lykken, 1998). Thus, it is not clear if the software can detect voice stress that is specific to telling lies.

Because most pseudoscientific and parapsychological topics are inherently interesting, they are particularly well suited for testing in methods courses (Lilienfeld, Lohr, & Motier, 2001). Students are likely to harbor pseudoscientific beliefs (Feder, 1985; Shermer, 1998), and instructors are eager to help them think more critically (e.g., Connor-Greene, 1993; Lawson, 1999, 2003). Indeed, Cole (1982) probably spoke for most instructors when he stated that the undergraduate curriculum should “seek to liberate the student from ignorance ... and from standing helpless before extravagant pseudoscientific assertions about human behavior” (p. 25).

Week 1 to Week 4: Teach Basic Concepts

Twenty-three sophomores and juniors who satisfied a statistics prerequisite spent the first 3 weeks learning the fundamentals of research methodology (e.g., quantitative and qualitative designs, operational definitions, confounding variables, reliability, validity, research ethics). To help students learn concepts quickly, the instructor selected a concise textbook (Martin, 2000) that addressed nearly all the methodological concepts identified in Boneau’s (1990) psychological-literacy survey.2 Near the end of the 4th week, students completed a comprehensive midterm examination.

During these initial weeks, students acquainted themselves with the literature on detecting deception. They searched electronic databases (e.g., PsycINFO, Criminal Justice Abstracts), creating annotated APA-style reference lists in the process. After students completed their electronic searches, the instructor allowed them to borrow articles from the archive he compiled previously. Although the instructor did not require students to read a minimum number of articles, he selected four specific items for all the students to read and summarize. For example, he selected Kubis’s (1973) research because it described a fascinating experiment involving a simulated theft.

Week 5: Design Study

After teaching themselves to operate Truster, students discussed how to test its effectiveness. Ultimately, they decided to design a simulated theft, similar to Kubis (1973). Specifically, they decided to randomly assign volunteers to be thieves or innocent suspects in a staged theft of two exams.

To help students identify threats to internal validity, the instructor noted several design flaws. For example, after discovering that innocent suspects would be ready for interrogation sooner than thieves, he asked, “How can we ensure that interrogators will remain blind to condition assignment?” Students provided solutions in class and via an online discussion board. To ensure efficiency, the instructor coordinated the creation of laboratory materials (e.g., consent form, interrogation script) based on student input and assistance.

Week 6 to Week 8: Conduct Study

All students worked in pairs to conduct the experiment. There were approximately 12 pairs of students and 63 participants who volunteered to participate for extra credit, which allowed each experimenter–interrogator pair to conduct approximately 5 sessions.

At the beginning of each session, the experimenter instructed thieves to steal the exams without being seen and to later deny their involvement. The experimenter informed control participants that there had been a staged theft and that they should simply deny their involvement. Each interrogation included a calibration procedure, a stimulation test, a guilty knowledge test, and several straightforward crime-

---

1Truster is available at http://www.truster.com for $108 and through various resellers for as low as $20. The software is easy to operate using a standard personal computer and microphone.

2The textbook contained 13 chapters, each approximately 20 pages long. The students read all but 3 chapters: dishonest scientific activities, multiple-variable designs, and statistical analyses.
related questions (e.g., “Did you steal the missing exams?”). Each session lasted approximately 35 min.

**Week 9: Analyze and Interpret Data**

Each student entered the complete set of data into SPSS and computed descriptive statistics and group comparisons for homework. Later in the week, students discussed how to interpret the results. Overall, Truster performed poorly, failing to distinguish between thieves and innocent suspects on all measures.

**Week 10: Finalize Research Reports**

Throughout the quarter, the instructor encouraged students to write specific sections of their APA-style research reports using the textbook as a guide. For example, after students reviewed the literature, the instructor directed them to write their introduction sections. This strategy enabled students to focus on each section independently, making an otherwise difficult writing assignment more manageable. Although the instructor provided feedback throughout the quarter, he reserved the 10th week for individual meetings. Completed reports were due during finals week.

**Results**

Overall, students performed well on their midterm examinations ($M = 87\%$, $SD = 9\%$) and on their research reports ($M = 84\%$, $SD = 10\%$), suggesting that the rapid pace of the course was not detrimental and that they succeeded in communicating their research findings effectively. Whereas we acknowledge that grades are inextricably linked to question difficulty and assessment criteria, these results suggest that students met course objectives. However, without comparative data (i.e., from a similar class taught using traditional methods), it is difficult to interpret the results because it is possible that students could have learned more (or less) if exposed to more traditional teaching techniques. Also, student evaluations provided evidence that they enjoyed the experience. On a scale that ranged from 1 (poor) to 5 (excellent), students rated the course favorably ($M = 4.55$, $SD = 0.67$).

**Discussion**

Some instructors may be leery of weaving courses around phenomena likely to produce null results. They fear students may become disappointed and be left with little to discuss in their reports. Our class experience suggests otherwise. After debunking Truster’s reported efficacy, students felt empowered and were eager to expose the software as a fraud. Nevertheless, the instructor noted that Type II errors are possible with all significance tests, and he highlighted the importance of future replication. Through these discussions, students learned that nonsignificant results are just as important as significant results, particularly when testing pseudoscientific claims. In their reports, students noted the utility of experimentation, and they wrote about important theoretical issues such as whether a unique physiological lie response exists.

Regardless, students who yearn for statistical significance can easily include factors likely to result in group differences. Instructors can find other interesting pseudoscientific topics to investigate in books (e.g., Hines, 2003; Ruscio, 2002), magazines (e.g., Skeptical Inquirer), and on myriad Web sites. For example, students can test suspicious product claims (e.g., “Can the Q-Ray ionized bracelet relieve pain?”), natural remedies (e.g., “Can herbal supplements improve memory?”), subliminal advertising (e.g., “Can subliminal primes influence behavior?”), extrasensory perception (e.g., “Can psychics perceive remote events?”), or even intercessory prayer (e.g., “Can the prayers of strangers have therapeutic effects on others?”). Whereas some topics lend themselves to laboratory investigations, others may require alternative procedures and slight adjustments to the proposed 10-week schedule.

**References**

Online and Writing: Teen Blogs As Mines of Adolescent Data

Elizabeth Mazur
Penn State University, McKeesport Campus

This article describes teenagers’ Internet Web logs or “blogs” and their potential and drawbacks as sources of firsthand material for the teaching of adolescence. Specifically, in a course on adolescence, undergraduates analyzed why adolescents write blogs; their function in terms of adolescent media use; the appeal and drawbacks of blogs for the writers and their readers; and the depiction of positive and negative emotions, friendship, and romantic relationships. Student survey responses over 2 semesters suggested increased knowledge of both the concerns of adolescents and of adolescent computer use as well as high satisfaction with use of teen blogs as an engaging and educational pedagogical technique.

According to recent media reports (Hill, 2003; Nussbaum, 2004; St. John, 2003), Web logs or “blogs” are journalism’s latest craze. News reports estimate that of the almost 6 million people in 2004 who were actively blogging, 51% were between ages 13 and 19 (Nussbaum, 2004; O’Connell, 2003). Because adolescent bloggers freely discuss their thoughts about numerous topics, I designed a learning activity for my course in adolescence that takes advantage of this rich source of primary data.

What Are Blogs and Why Should Students Analyze Them?

Blogs are basically online journals, pages of commentaries and photos that their authors often update monthly, weekly, or daily, sometimes moments before you read them. Although many teen blogs are brief experiments, for numerous adolescents they become a paradoxical way of life, a public record of one’s private thoughts.

Although teachers of undergraduates have used the Internet successfully to teach about psychological disorders (Casteel, 2003) and developmental psychology (Mazur, 2003; Sheldon, 2004), no one has discussed exploiting blogs as firsthand source material for teaching about adolescence. Similar to the taped reports of National Public Radio’s (NPR’s) Teenage Diaries (described in Sheldon, 2004) and contrary to typical Web pages, blogs “have voice and personality. They’re human” (Grossman, 2004, p. 66). However, unlike the 15 Teenage Diaries currently available, the pool of blogs is vast and unedited by adults.

Learning Activity Goals and Description

Activity Goals

I had three major goals for the blog learning activity. First, students examined blogs as primary sources of data on adolescent friendship, romantic relationships, and the attendant emotions. Second, students were to learn of blogs’ existence as part of current teen culture and Internet use. Only 11% of Internet users visit blogs written by others (Lenhart, Horrigan, & Fallows, 2004). Finally, students analyzed blogs as a specific type of media, as described by Santrock (2003).

Description of Teen Blogs

I have used this activity twice in small 400-level adolescence courses that met for lengthy blocks of time. During a 1.5- to 2 hr period in a reserved computer room, students analyzed two to three blogs individually or in pairs either by doing a search (e.g., “Teen + blog”) or by visiting hosting Web sites (e.g., DeadJournal.com, Diaryland.com, Freeopendiary.com, Livejournal.com, My-diary.org, Teenblogs.studentcenter.org).

Although formats vary only slightly, the sites that host the blogs may attract somewhat different crowds (Nussbaum, 2004); DeadJournal.com, for example, has a somewhat morbid theme. Although many posts are just a few lines, the blogs...
Enhancing and Assessing Critical Thinking in a Psychological Research Methods Course

Emily Stark

Teaching of Psychology 2012 39: 107
DOI: 10.1177/0098628312437725

The online version of this article can be found at:
http://top.sagepub.com/content/39/2/107

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Society for the Teaching of Psychology

Additional services and information for Teaching of Psychology can be found at:

Email Alerts: http://top.sagepub.com/cgi/alerts
Subscriptions: http://top.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Mar 20, 2012
What is This?
Enhancing and Assessing Critical Thinking in a Psychological Research Methods Course

Emily Stark

Abstract
This article presents an assessment of critical thinking skills and paranormal beliefs in students taking a research methods course. The course emphasized separating scientific claims from pseudoscientific assumptions and used real-world examples to teach students to think critically; however, the course did not use examples of the paranormal to teach about pseudoscience. There was a significant decrease in students' paranormal beliefs at the end of the course and significant improvement in their abilities to evaluate psychological research; however, there was not a significant change in their general critical thinking abilities. In addition, all measures related to course performance, such that students who showed lower levels of paranormal beliefs and greater general critical thinking abilities at the outset of the course earned more total points. The author discusses how the course builds skills recommended for psychology majors and identifies areas for future research in scientific thinking and belief in pseudoscience.

Keywords
critical thinking, research methods, pseudoscience

People are constantly bombarded with advertisements and news stories telling us what to eat, how to exercise, what music to play for our infants to make them smarter, which products will save the environment, and which politicians to vote for. How do people make sense of this information and decide what to believe? The ability to critically evaluate the countless stories and advertisements we see and hear every day is important and necessary if we wish to make wise decisions about our health, our money, and our happiness. However, in the United States, there are signs that critical thinking is not a universal skill. The National Science Foundation expressed concern that belief in and use of alternative medicine practices, such as magnetic healing, homeopathy, or herbal medicine, is growing, despite a lack of scientific evidence that these therapies or interventions actually work (National Science Foundation, 2002). Also, a Gallup survey conducted in 2005 found that 73% of Americans held at least one paranormal belief, such as belief in extrasensory perception or that houses can be haunted. In a climate that seems to value belief over skepticism, how can instructors teach abilities to think critically about information and to distinguish science from pseudoscience?

The American Psychological Association (APA, 2007) Guidelines for the Undergraduate Psychology Major list “critical thinking skills” as a learning goal for undergraduate students, specifically stating, “Students will respect and use critical and creative thinking, skeptical inquiry, and, when possible, the scientific approach to solve problems related to behavior and mental processes” (p. 10). This connection between understanding the scientific approach to solving problems and developing critical thinking skills suggests that a course on research methods and design would offer a prime opportunity to teach students to critically evaluate information and further develop critical thinking abilities. In addition, students who are not planning on continuing to graduate school may not immediately see the value in learning about research methods, not expecting that they themselves will ever be designing studies or conducting research. Situating a research methods class into a broader goal of teaching critical thinking makes the course more relevant to all students, regardless of their future career plans.

Research on critical thinking and building these skills in the domain of psychology courses is growing rapidly (see Dunn, Halonen, & Smith, 2008, for a compendium of critical thinking exercises and teaching strategies). Halpern (1998) describes a skills-based approach to teaching critical thinking in students and emphasizes the importance of using real-world contexts.
to build these skills to allow students to apply their critical thinking abilities in other domains. Students need to appreciate the underlying structure of problems as they learn to reason about them to be able to transfer their critical thinking skills to other problems that they encounter in their lives. Some instructors have worked to build these abilities through incorporating topics related to pseudoscience in their classes (e.g., Morier & Keeports, 1994; Wesp & Montgomery, 1998). Being able to identify a claim as pseudoscientific is a skill that could be applied to many different problems, which facilitates the transfer of learning.

Several studies have focused on the extent to which students transfer their knowledge of science to evaluating pseudoscientific or paranormal claims. For example, Manza et al. (2010) examined the effects of including specific paranormal examples to introduce students to pseudoscience in an undergraduate quantitative methods course. In this study, students’ beliefs in the paranormal only decreased after taking the course that included specific paranormal examples; simply learning about quantitative methods was not enough to increase skepticism about the paranormal. The authors suggest that students with higher levels of critical thinking ability or scientific reasoning skills may approach pseudoscientific claims with more skepticism than students with lesser critical thinking abilities such that critical thinking skills may lay a foundation for later skepticism. Students with less extensive critical thinking skills may need direct examples of debunking paranormal claims to influence their beliefs; students with more extensive critical thinking skills may be better able to transfer what they learn about science to evaluating their own beliefs.

McLean and Miller (2010) examined this potential connection between critical thinking abilities and skepticism about the paranormal through examining students’ improvement in critical thinking skills and change in paranormal beliefs after taking a course designed to help students distinguish between science and pseudoscience. The course presented topics on pseudoscience and superstition, along with readings on skepticism and critical thinking. Students taking this course showed a decrease in their paranormal beliefs after completing the course and an improvement in their critical thinking skills. However, similar to Manza et al. (2010), students enrolled in a research methods course that did not incorporate the topics of identifying pseudoscience did not show any change in their paranormal beliefs, although they did improve their critical thinking abilities. McLean and Miller suggest that direct instruction in parapsychology and pseudoscience might be necessary to influence students’ paranormal beliefs. However, in both Manza et al. and McLean and Miller, the “non-paranormal” course used for comparison did not address the elements of pseudoscience. The comparison courses focused on lower-level statistics (Manza et al., 2010) or on statistical concepts and research methods (McLean & Miller, 2010). It may be too big of a leap for students to move from thinking about statistics to evaluating the paranormal, even if critical thinking abilities are relevant to each area. However, students who are given some instruction in elements of pseudoscience may be able to apply this understanding to thinking about the paranormal, even in classes that do not rely on including specific paranormal topics.

The current research presents an assessment of changes in paranormal beliefs and critical thinking skills as a result of completing a lower-level course on research methods in psychology. The current study extends prior research through examining the effects of a course that does focus on the elements of science compared to pseudoscience but does not focus on the paranormal as a core topic of the class. This raises the question of whether just learning about science compared to pseudoscience is enough for students to realize the pseudoscientific elements of the paranormal and effectively reduce their belief in the paranormal. Or, do students need explicit instruction in paranormal topics and research to influence their beliefs?

To determine if this research methods course influenced students’ critical thinking abilities and their beliefs in the paranormal, I used multiple assessments, including paranormal belief measures and two critical thinking ability tests, at both the beginning and end of the semester. I expect that students taking the research methods course will show an increase in their critical thinking skills over time. Also, even though the course did not directly address paranormal beliefs, I expect that, because of the focus on teaching pseudoscience, paranormal beliefs will decrease as compared to the pretest at the beginning of the course. Finally, I examined how the critical thinking and paranormal beliefs measures related to students’ course performance. I expect that students’ scores on the critical thinking measures will positively relate to their overall course performance, representing successful learning of course concepts and an ability to apply those concepts when tested on course exams.

Method

Participants

Participants consisted of 78 undergraduate students (24 male, 54 female), all of whom were enrolled in a research methods in psychology course. They completed the study measures as part of course activities at both the beginning and end of the course. The majority of the entire sample was Caucasian (98%).

Course Information and Relevant Assignments

The course examined in the current study is a lower-level research methods course that is required of all psychology majors. Students need to complete a statistics course before enrolling in research methods, and they most often enroll in research methods in their sophomore or early junior years of college. The course primarily focuses on teaching different types of research design and issues related to conducting research, analyzing data, and writing APA-style reports of research. Students develop research projects, collect and analyze data, and write up their results.

To teach about research designs, students were introduced to elements of pseudoscience, learned how to separate scientific claims from pseudoscientific assumptions, and were asked to evaluate media presentations of research, which all represent...
broader critical thinking skills. Students completed a total of five assignments, described briefly below, that focused on identifying pseudoscience and critiquing scientific designs. All of these assignments were completed individually by students outside of class and then discussed as a class as part of lecture activities.

1. Students found an advertisement that incorporated elements of pseudoscience and to describe those elements and how they were present in that advertisement.
2. Students found a newspaper article presenting a research study, to identify the design used, and to identify any flaws or gaps in how the research was described in the newspaper article.
3. Students found and reviewed magazine or newspaper articles about the research regarding possible connections between cell phone use and cancer and identified misleading statements about the research design and conclusions in the media articles.
4. Students found and reviewed magazine or newspaper articles about research showing no connection between diagnoses of autism and vaccination and identified flaws in reasoning and misleading statements about the research and other perspectives presented in the media articles.
5. Students viewed a documentary and news clips about the use of facilitated communication with coma patients and autistic children (a method that has been shown to not be accurate or valid). Students identified flaws in reasoning in people supportive of this method and developed experimental designs to test this method.

In addition to these assignments, students were often given examples of research studies and were asked to identify potential threats to internal validity or other flaws in the research as part of lecture and in-class activities. Also, in their write-ups of their own experiments, they were asked to identify flaws or limitations in their own studies and to propose how future researchers might design a different study to test their ideas. Over the semester, students practiced designing and critiquing research in multiple contexts as well as drawing conclusions from different types of studies. These course aspects reflect the recommendations posed in Bensley (2010) for teaching and assessing critical thinking in psychology, including infusing critical thinking into course content, giving students practice in thinking critically about psychological studies, and fitting the critical thinking content into the overall goals of the course. Also, it is important to note that although a central focus of this study was on the evaluation of the students’ critical thinking skills, the course aimed to increase students’ awareness of pseudoscience and to describe those elements and how they were present in that advertisement.

Materials

First, participants completed the Revised Paranormal Beliefs Scale (Tobacyk & Milford, 1983). This scale consists of 25 items and includes seven subscales of traditional religious beliefs, belief in psi abilities such as mind reading, belief in reincarnation, belief in extraordinary life forms such as the Loch Ness Monster, and belief in precognition, such that dreams can tell the future. These measures were rated on scales of 1 to 5, with 5 indicating greater levels of belief. The overall alpha for this scale in this study was .85.

Next, participants completed the Psychological Critical Thinking Exam (PCTE; Lawson, 1999). This test asks participants to evaluate claims presented in scenarios. Each scenario presents a problematic claim, such as drawing a broad conclusion from a limited sample. Participants are asked to determine if there is a problem with the claim presented in each scenario and, if so, to describe the problem. Their responses were graded using the following scale: 0 points if they did not identify a problem, 1 point if they identified there was a problem but were unable to explain it, and 2 points if they identified the major problem with the scenario. Participants viewed seven total scenarios, so scores could range from 0 to 14, with higher numbers indicating better critical analysis of the claims. (The full PCTE includes 14 total scenarios; I randomly selected half to use for this study because of time constraints in class.)

In addition, participants completed the Cornell Critical Thinking Test, Level Z (CCTT; Ennis, Millman, & Tomko, 2005). The CCTT is a 52-item, multiple-choice, critical thinking test designed for college and graduate students. It tests the skills of induction, deduction, making inferences from observations, assessing credibility, and identifying assumptions.

Procedure

Students in the research methods course completed the experimental measures as part of course activities within the first 2 days of the course and again within the final 2 days of the course. All participants completed the measures in the same order: Paranormal Beliefs Scale, PCTE, and CCTT.

Also, participants’ scores on the measures were related to their overall point totals for the course. For this course, students could earn a maximum of 360 points, with 50% of the points coming from exams, 25% coming from assignments including the five focused on pseudoscience described above, and the final 25% coming from papers about the students’ own experiments.

Results

When comparing the scores of the students from the beginning of the course to the end of the course, there were significant changes in students’ scores on many of the measures. Students had a significantly lower score on the Paranormal Beliefs Scale at the completion of the class ($M = 2.26$) compared to the beginning of the class ($M = 2.51$), $F(1, 56) = 23.24, p < .001, \eta^2 = .29$. When examining the specific paranormal beliefs subscales, four showed significant decreases in score. The analyses of the subscales included a Bonferroni correction.
to adjust for the multiple analyses, setting the alpha value for significance for these analyses at .008. The students had a significantly lower score on beliefs in psi abilities at the completion of the course (M = 1.94) compared to the beginning of the class (M = 2.16), F(1, 63) = 7.62, p < .008, η² = .11. Students also showed a significant decrease in beliefs in witchcraft, with scores at the end of class (M = 1.84) lower than scores at the beginning of the class (M = 2.20), F(1, 60) = 19.55, p < .001, η² = .25. Belief in spiritualism also decreased, with scores at the end of class (M = 2.17) lower than scores at the beginning of class (M = 2.55), F(1, 63) = 17.03, p < .001, η² = .21. Finally, belief in precognition also decreased, with students again showing a significantly lower score at the end of the course (M = 2.95) compared to the beginning (M = 3.25), F(1, 62) = 8.20, p < .006, η² = .12. (See Table 1 for the pretest and posttest scores with standard deviations for all measures.)

Also, students showed a significant increase in their scores on the PCTE from the beginning of the course (M = 6.62) to the end of the course (M = 9.60), F(1, 62) = 72.32, p < .001, η² = .54. However, there was not a significant change in students’ scores on the CCTT from the beginning of the course (M = 27.08) to the end of the course (M = 26.90), F(1, 51) = 0.06, p < .81, η² = .001.

There were very few relationships between the critical thinking measures and the Paranormal Beliefs Scale. The only exception is that students who scored higher on the CCTT at the beginning of the course showed lower levels of superstitious belief at both the pretest and the posttest (rs = −.30, −.25, p < .04). However, after including the Bonferroni correction for these analyses with the seven subscales of the paranormal beliefs measure, the significance of these correlations did not reach the adjusted alpha level of .008, such that there were no significant relations between either critical thinking measure and the Paranormal Beliefs Scale. Also, there was not a significant relationship between the two critical thinking measures at the pretest (p > .21), but there was a significant, positive correlation between the PCTE and the CCTT at the posttest (r = .46, p < .0001).

To determine whether participants’ scores on the study measures predicted their performance in the research methods class, I regressed the total course points that students earned on their scores on the PCTE, the CCTT, and their overall Paranormal Beliefs Scale score, from both pretest and the posttest. First, I examined whether any of the pre-test measures predicted students’ final performance. Both the CCTT score (B = .306, p < .04) and the overall Paranormal Beliefs score (B = −.343, p < .02) from the beginning of the course related to students’ course performance, such that students who scored higher on the CCTT and lower on the Paranormal Beliefs measure earned higher course grades. However, when examining connections between performance on the study measures at the end of the course and course grades, only the PCTE score (B = .386, p < .01) predicted students’ total course points, such that students who had higher scores on this critical thinking measure at the end of the course also earned more points in the course.

Discussion

The research described here adds to the growing literature on critical thinking through incorporating multiple measurements of critical thinking and belief in the paranormal. This study used a pretest–posttest design to test the acquisition and development of critical thinking skills in students taking a psychology course in research methods, which emphasized learning the elements of science versus pseudoscience. Also, this research examines relationships between these measures of critical thinking and students’ overall performance in the course, an aspect that distinguishes the current study from much of the previous research.

Students in the research methods course showed a significant decrease in paranormal beliefs overall from the beginning to the end of the course, with significant decreases also on several of the subscales. This shows that even though the course did not directly include content related to the paranormal, students were able to transfer their learning of pseudoscience to reflect on, and decrease, their belief in the paranormal, contrary to the findings of McLean and Miller (2010) and Manza et al. (2010). This highlights the importance of emphasizing pseudoscience and shows that students are able to transfer what they learn in their classes to evaluating their own beliefs.

It is important to note that most of the scale means for the Paranormal Beliefs Scale are below the midpoint, indicating that this group of students is fairly skeptical about the paranormal, even at the pretest. Although measures of paranormal beliefs have been commonly used to assess the level of skepticism in participants, perhaps other measures of acceptance of pseudoscience would be more useful, particularly for groups of participants who show low levels of belief in the paranormal. For example, future researchers could develop a test of the ability to identify elements of pseudoscience in the context of critiquing misleading advertisements or reports of alternative medicine. This might be a more relevant way to identify differences in students’ ability to think scientifically, compared to...
just focusing on the paranormal. Johnson and Pigliucci (2004) as well as Walker and Hoekstra (2002) used a measure of pseudoscientific beliefs that included paranormal beliefs as well as beliefs in New Age medicine (such as magnet healing) and beliefs in aliens and found no significant relationship between participants’ knowledge of science and their levels of pseudoscientific beliefs. Also, the lowest levels of skepticism were seen for the question about magnet healing, suggesting that even students who did not believe in the paranormal endorsed this New Age medicinal claim. Research that more closely examines this aspect of pseudoscience, of evaluating claims related to healing or health, will add to our understanding of how to build critical thinking skills and give students the tools to evaluate their own beliefs.

In the current study, there were no relations between the critical thinking measures and participants’ paranormal beliefs. However, participants with low levels of paranormal belief overall when beginning the course scored more course points overall compared to those with higher levels of paranormal belief, suggesting an inverse relationship between belief in the paranormal and students’ ability (or willingness) to think scientifically, as measured by performance in this research methods course. This is similar to the results of Messer and Griggs (1989), where students who had higher levels of belief in the paranormal also had lower course grades in an introduction to psychology course. Students with higher levels of paranormal beliefs may be more dismissive of scientific methods in general, compared to students with lower levels of paranormal belief. This could have hindered learning in the students with higher levels of paranoid beliefs, leading to poorer performance in the course in the current study. Or, high levels of paranormal belief may indicate people who tend to think at a surface level about concepts rather than expending the time and energy to analyze situations in-depth using scientific approaches. These people may be perfectly capable of thinking logically or scientifically in areas that apply to them, or when they are motivated to analyze a certain topic, but may otherwise not take the time to analyze and examine evidence, defaulting to a general position of belief in most topics.

This could also explain the lack of a correlation between the two critical thinking tests, the PCTE and the CCTT, at the outset of the course. When presented with the psychological scenarios of the PCTE, the students may have thought of it as relevant and specific to the course, focusing their attention on the situations. However, when they moved on to complete the CCTT, the lack of any connection to psychology or evaluating experiments in the test questions may have led some students to rely less on their analytical thinking abilities and just take a surface approach to answering the questions, such that their score on this test no longer accurately assessed their ability to think critically. Future research should develop measures of general acceptance of pseudoscience and of willingness to think scientifically to better identify how this ability relates to learning in general and specifically in science courses. Also, individual difference measures such as the Need for Cognition (Cacioppo & Petty, 1982) could be used to try to identify students who are less likely to think critically to try to tease apart differences in test scores or paranormal beliefs resulting from an ability to think critically and willingness to spend the time and energy to think critically (also see Bensley & Murtagh, 2012, for other examples of measures to assess critical thinking dispositions and attitudes).

When examining performance on the critical thinking tests, the participants showed a significant improvement in their scores on the PCTE, evidence that participation in the research methods course did improve their abilities to critically evaluate psychological claims. Also, students who scored high on this measure at the posttest earned more total points in the course than did students who had lower scores on the PCTE, showing a close correspondence between this critical thinking measure and the overall focus of the course. This measure asks participants to evaluate experiments and identify flaws in designs, which was a major focus of the course itself.

However, there was no significant difference in students’ scores on the CCTT, although performance on the CCTT at the beginning of the course predicted overall course performance. The CCTT focuses on the more general skills of making deductions and inferences from observations and identifying assumptions. Although these are aspects of understanding and critiquing psychological experiments that were addressed in this course, the CCTT may have been too different of a context for these skills, and students perhaps were unable to transfer what they learned in the course to their performance on this test. Bensley, Crowe, and Bernhardt (2010) also found limitations in students’ abilities to transfer critical thinking to other domains; these authors examined the specific critical thinking skill of argument analysis in students who had taken a course that included explicit critical thinking instruction and students in a general research methods course, without a specific focus on critical thinking. Only students in the course with explicit critical thinking skill instruction showed increases in their ability to analyze arguments, showing that just learning about research methods was not enough to build this critical skill in students. This is certainly an area for future research on teaching and learning: How can we build students’ ability to transfer skills to new contexts and to understand how to solve problems different from those they have seen before? In the current study, the focus of the research methods course on psychological experiments seems to have made it difficult for the students to apply their critical thinking abilities when faced with the more general, nonpsychological questions of the CCTT. This identifies the limitations of students’ abilities to transfer course content into other domains: Although learning about pseudoscience did influence their skepticism about the paranormal, this did not transfer to the more general critical thinking abilities measured by the CCTT.

It is important to determine if general research methods courses like the one assessed in the current study effectively improve critical thinking skills because these courses are a common presence in undergraduate psychology programs. Not all programs offer specialized courses on pseudoscience or paranormal topics, but most, if not all, psychology programs require their majors to take a research methods course. Are
students able to generalize from learning the aspects of pseudoscience to broaden their critical thinking skills outside of the specific course content? Will learning about pseudoscience influence students’ paranormal beliefs as well as their critical thinking skills? The results of the current study suggest that taking a research methods course that incorporated information about distinguishing science from pseudoscience resulted in a lower acceptance of paranormal beliefs and improved abilities to evaluate psychological assumptions.

Future work in this area should focus on how to improve students’ abilities to transfer skills learned in a domain-specific course (in this case research in psychology) to other situations where critical thinking is relevant to add to the growing literature on critical thinking and building these skills. In addition, it is important for future researchers to continue to examine the relations between students’ acceptance of pseudoscience and the ability to learn about science. This can lead to identification of mind-sets that may facilitate, or block, learning of more general critical thinking skills and the ability to transfer those skills to new domains. As mind-sets that promote or hinder learning are identified, researchers and instructors can then develop and test course activities that promote critical thinking in all students.

Declaration of Conflicting Interests
The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by a Teaching Scholar Fellowship provided to the research, authorship, and/or publication of this article: This research.

References


A letter-writing exercise renders concepts from statistics and research methods more meaningful by allowing students to put complex ideas into their own words. Students select topics and write about them in letters to peers in another section of the same course. Most students are able to correctly explain abstract course material to peers in the form of a letter.

Most courses in the undergraduate psychology curriculum generate student enthusiasm and understanding with ease. Exceptions include statistics and research methods, where applying ideas from these courses to the interpretation of human behavior can seem foreign, not intuitive. In addition to the difficulty of the material presented in these courses, there is a problem with its novelty. In my experience, conceptual understanding in statistics or research methods is often sacrificed for rote learning—students learn statistical procedures (e.g., the t test) or definitions (e.g., random sampling) but draw little meaning from them.

I use a letter exchange exercise when teaching statistics and research methods to make the course material meaningful and to promote its effective communication. Students select topics from statistics or research methods, then write to peers in another section of the same class. In turn, these peers respond with letters of their own. Putting complex ideas into
Peer Writing, Letters, and Essay Exchanges

Sharing writing with student peers is increasingly commonplace in English courses using workshop approaches to teaching composition (e.g., Elbow & Belanoff, 1995). Workshop courses involve writing exercises inside and outside the classroom and regular opportunities for reading and commenting on peer efforts. Although less common in psychology courses, forays using peer review (Dunn, 1994; Haaga, 1993) and peer collaboration (Dunn, 1996; Dunn & Toedter, 1991; Goldstein, 1993) do exist. Books adopting a workshop focus for the social sciences, too, are available (e.g., Williams & Brydon-Miller, 1997). Letter writing is not common in psychology courses, although in one exercise, Dunn (1992) had students write letters in response to the famous “Why War?” correspondence between Albert Einstein and Sigmund Freud (1932/1964). Students examined this historical exchange of letters on human aggression and responded to the authors in a personal way.

Having students exchange short essays with one another combines the virtues of peer collaboration with the personal appeal inherent in letter writing. Konzem and Baker (1996) used an essay exchange exercise to teach students in finance and physiology to effectively communicate their respective course materials to others. This essay exchange required student writers to do three things: develop and exchange an essay with a peer, review the peer’s comments on it, and clarify any points raised therein. Konzem and Baker found that discourse on the course materials improved, as did the quality of student writing.

Writing About and Responding to Statistics and Research Methods

I gave students in two sections of my statistics and research methods course an exercise similar to that of Konzem and Baker (1996), but I asked them to write individualized letters instead of essays. Students received the name of a peer in the other section of the class to write to and received the following instructions:

Your goal is to write a letter in which you explain one thing you have learned about statistics (research methods) that you find interesting, compelling, confusing, or otherwise noteworthy. Tell why it was important to have learned it, and explain why it is important for students of psychology to know it. Finally, discuss whether students outside of psychology would benefit from this knowledge, as well.

Before midterm, students wrote on a statistical topic; after midterm, research methodology. They could send printed letters to peers or use e-mail. Each student writer received a letter from a peer and responded in writing within a week. After reading the response letter, students could write another

letter clarifying any points as necessary. I received copies of all student letters.

I read all the correspondence. Letter content varied, but students obviously took the exercise seriously. No one selected a topic that was too easy, and the letters were not too short—the average length was approximately two double-spaced pages. Topics from statistics included frequency distributions, the normal curve, percentile ranks, measures of central tendency and dispersion, correlation, regression, analysis of variance, post-hoc comparisons of means, and graphing data. The research methods letters examined issues such as naturalistic observation, the scientific method, the Nuremberg Code, informed consent, ethics of experimentation, demand characteristics, factorial designs, operational definitions, internal validity, random sampling procedures, and single-subject research designs.

Students appreciated the opportunity to show peers what they had learned. In the course evaluations, one student wrote, “I thought the letters were a good way for us to see how much we were learning.” Another remarked that, “The letter writing project was fun, and it made you understand the topic … when you read it in print, it was quite obvious when it was weak. (I did my letters over a few times).” Having the students put their ideas in writing may have rendered many of the more abstract ideas from statistics and methods much less threatening. I also found that writing letters encouraged students to use everyday examples to demonstrate or clarify concepts. One student, for example, explained regression to the mean by discussing the various heights of family members from different generations.

I evaluated three dimensions of the students’ letters: topic presentation, quality of writing in the letters, and the peer component. Ideally, students should have explained abstract concepts from methods or statistics by defining them clearly and then providing concrete examples to support their definitions. Using these two criteria, I globally evaluated each initial statistics or methods letter as being factually “correct” or “incorrect” in presenting a topic. I evaluated 83% of the statistics letters (N = 36) and 89% of the method letters (N = 37) as factually correct. When a letter’s content was incorrect or otherwise difficult to interpret, peer reviewers often corrected the error or suggested alternative ways to present the topic.

Perhaps because it is the most objective part of the exercise, grading the grammar, punctuation, and spelling within the letters was not difficult. Evaluating writing style was somewhat more difficult, and instructors will no doubt apply their own standards of what constitutes good prose. The peer component was easy to assess, as instructors need only verify that original letters were sent and that peer responses came back (only about one third of the letter writers believed that sending a clarification note to peers was necessary).

There are alternative ways to conduct this exercise. Peer writers need not be from two different sections of the same class—they could be in the same class, an altogether different psychology course or, following Konzem and Baker (1996), from a different discipline. A campus computer network could make the letter writing still more interactive; faculty members who are so inclined could develop a chat group devoted to issues in methods and statistics.
Conclusions

In my view, exchanging letters can promote effective communication of conceptually challenging material from statistics and research methods (see also Beins, 1993). Student writers apply this material to real-world examples while, as peers, they share abstract ideas and constructively critique others’ work. Writing ties these activities together by making the novel more familiar, even personal. To modify an old teaching maxim, the best way to understand something is to write about it.

References


Notes

1. I am grateful to the students in my statistics and research methods classes for allowing me to use their letters. Robert Brill, Randolph Smith, and three reviewers provided helpful comments on an earlier version of this article.

2. Send correspondence to Dana S. Dunn, Department of Psychology, Moravian College, 1200 Main Street, Bethlehem, PA 18018–6650; e-mail: dunn@moravian.edu.

Components of the Reader’s Guide

I outline the structural components students put into a reader’s guide and give examples of the components from guides students in undergraduate child psychology and masters-level general psychology courses have prepared. I give students a handout that lists major components their reader’s guide should have.

Content Outline

The sections of the content outline vary by topic but may include historical background, theoretical and methodological basics, major ideas or research areas, and problems or issues that are important to the area. An outline for a graduate student’s guide on memory included history, brain structures,
Conclusions

In my view, exchanging letters can promote effective communication of conceptually challenging material from statistics and research methods (see also Beins, 1993). Student writers apply this material to real-world examples while, as peers, they share abstract ideas and constructively critique others’ work. Writing ties these activities together by making the novel more familiar, even personal. To modify an old teaching maxim, the best way to understand something is to write about it.

References


Notes

1. I am grateful to the students in my statistics and research methods classes for allowing me to use their letters. Robert Brill, Randolph Smith, and three reviewers provided helpful comments on an earlier version of this article.

2. Send correspondence to Dana S. Dunn, Department of Psychology, Moravian College, 1200 Main Street, Bethlehem, PA 18018–6650; e-mail: dunn@moravian.edu.

The Reader’s Guide As an Integrative Writing Experience

Bruce B. Henderson
Western Carolina University

In this article, I describe a project that helps students get an overview of the literature on a topic in psychology and develop a mechanism for conveying that view to others. A reader’s guide outlines a topic; provides brief descriptions of major concepts and contributors; identifies “hot topics”; annotates important related books and articles; and lists major journals, books, and Internet resources. In addition to providing conceptual overviews, reader’s guides have been useful as study guides and as aids to teaching.

Teachers of psychology have developed a wide array of writing assignments to engage student learning and thinking (Boice, 1982; McGovern & Hogshead, 1990). Some writing assignments, such as journals, emphasize making material more personally meaningful (e.g., Dunn, 1994). More commonly, assignments focus on developing professional skills in presenting empirical research or research reviews (e.g., Froese, Gantz, & Henry, 1998; Snodgrass, 1985). Vipond (1993) argued that psychology teachers need to expand the genres and styles they typically require students to use in writing assignments. Vipond also argued that it is particularly important for students to write for audiences other than their instructors.

In this article, I describe a writing project I have found useful for helping students get a conceptual overview of a subspecialty of psychology while writing for an audience of their peers. The writing assignment is called a “Reader’s Guide to the Literature in X.” X can be as big or as small a unit as you want to make it, as large as perception or as small as the moon illusion. In preparing a reader’s guide, students learn about the scope of the theory and empirical research available on a topic. Students also learn to write for readers who might use the information reported in the guide rather than for an instructor who judges a writing performance.

Components of the Reader’s Guide

I outline the structural components students put into a reader’s guide and give examples of the components from guides students in undergraduate child psychology and masters-level general psychology courses have prepared. I give students a handout that lists five major components their reader’s guide should have.

Content Outline

The sections of the content outline vary by topic but may include historical background, theoretical and methodological basics, major ideas or research areas, and problems or issues that are important to the area. An outline for a graduate student’s guide on memory included history, brain structures,
means of measuring memory, types of memory, memory processes, theories of memory, and memory disorders. Students provide multiple annotated references on each topic with attention to classical contributions and emphasis on currency. Annotation and currency are important to ensure that students are using a wide variety of sources.

Theorists and Contributors

In this section, students list the dozen or so most important theorists and researchers in the area with two or three sentences identifying the persons and why they are important. An undergraduate student’s reader’s guide for attachment listed the pioneers Freud, Bowlby, and Ainsworth, and more recent contributors, including Belsky, Main, and Waters.

Central Concepts

Students make a list of and describe the 10 to 12 most important concepts or phenomena in the area. I tell students to list concepts that almost anyone should know (of course, what that means depends on the level of the students in the class). Concepts listed in an undergraduate’s guide on infant cognitive development included assimilation and accommodation, habituation, stage, deferred imitation, and object permanence.

Hot Topics

The hot topics section of the reader’s guide is the heart of the project and the hardest part to do. Students peruse the most recent books, chapters, and journals to see which topics appear repeatedly or seem to have engendered debate. Up-to-date citations are an essential requirement of the hot topics component. Students talk with the department’s faculty members who are most likely to know what the major issues are and can speculate on what is going to happen in the future. Sometimes the hot topics are related to old questions. For example, in a graduate student’s guide on intelligence, a student listed the nature–nurture issue. In a guide on thinking, a student listed the question of consciousness and described recent work on brain research and artificial intelligence. Other hot topics may be listed because they have gained attention in the popular media (e.g., eyewitness testimony in a guide on memory), because they have led to new areas of research (e.g., behavior genetics research in personality or theory of mind applications to autism), or because of their practical applications (e.g., performance assessment in tests and measurements).

Major Resources

Students include in this section major handbooks and journals in the area and identify major Internet resources. For example, a guide on thinking and problem solving listed recent books; chapters in the Annual Review of Psychology; major journals such as Cognitive Psychology, Memory and Cognition, and the Journal of Educational Psychology; and Internet resources with links to online chats, online journals, and sites that provide practice in problem solving.

Role of the Reader’s Guide in Courses

I assign students topics after considering their preferences. Students then prepare guides, often in consultation with faculty members who have the appropriate expertise. At times I have had small groups of students collaborate in developing guides. Students or groups provide class members with copies of their guides and lead a class discussion of the hot topics. There are several logistical issues to consider in using reader’s guides. First, instructors need to provide considerable assistance, particularly in identifying hot topics. Undergraduates usually need more assistance than graduate students in identifying hot topics. I urge students to use resources that are particularly helpful in identifying hot topics such as Contemporary Psychology, the comment section of the American Psychologist, APA Monitor, Annual Review of Psychology, and Psychological Review. Second, grading standards have to be considered. I have graded reader’s guides holistically but with an emphasis on currency of annotated citations and hot topics. One guiding question I have used with graduate students is: “Would a student using this guide to prepare for a comprehensive exam on this topic find it useful?” Reader’s guides are not very useful for purposes of discrimination because most students do quite well on the assignment. Finally, the amount of time students spend on a reader’s guide depends on a number of variables, including the complexity of the topic, the number of entries and citations in each section required by the instructor, and, perhaps most important, how many hot topics the student needs to identify and characterize.

Student Reactions

Students’ reactions to the reader’s guide assignment generally have been quite positive. In a recent graduate class (n = 9), on a scale of 1 (strongly disagree) to 6 (strongly agree), all rated the statement “I learned a good deal through the process of developing my Reader’s Guides” with a score of 6. The average rating in response to the statement “The development of the Reader’s Guide was a challenging assignment” was a 5.8 (SD = 0.04). Undergraduate and graduate students report liking the opportunity to get the “big picture” and having a sense of what is most important in the most current literature (what textbooks often fail to convey). Students frequently report using the guides they have collected in preparing for the psychology subject test of the Graduate Record Examination or in preparing class lectures when they teach. The major limitations on the reader’s guide assignment are time and availability of library resources. However, the active learning characteristics and scholarly timeliness of the reader’s guides appear to make them worthwhile.
A One-Minute "Intelligence" Test

Richard A. Griggs
University of Florida

Testing, especially abilities testing, is an essential topic in psychology education because of its importance to students in both academics and the nonacademic world. In this article, I describe a class activity requiring minimal preparation that serves as an engaging stimulus for discussion of many important aspects of testing. Although I describe my use of the activity in introductory psychology, it has other uses in this course and in other psychology courses involving testing.

College students are clearly familiar with testing, both academic and standardized (e.g., ACT, SAT) tests. Similarly, they realize the importance of test performance to their success in academics and to their future careers. Given its importance, testing is a constant topic for public debate (e.g., the recent "Bell Curve" controversy; Herrnstein & Murray, 1994). Thus, it is essential for psychology teachers to provide students with a solid understanding of testing, especially abilities testing. A brief examination of courses and textbooks indicates that psychology students realize the importance of these topics for their students. Most schools have courses in intelligence; Messer, Griggs, and Jackson (1999) found that 82% of the schools surveyed offered such a course. In addition, intelligence and testing are topics covered in several other courses beginning with the introductory course. It is crucial to do a good job on these topics in the introductory course because this is the only experience most undergraduates have with psychology. I checked A Compendium of Introductory Psychology Textbooks (Jackson, Griggs, Marek, & Christopher, 1998) and found that all but one of the current full-length introductory texts have a chapter or major chapter section on intelligence and testing, providing further evidence of its importance in psychology. I designed the activity I describe in this article to open the unit in my introductory course on this topic.

In developing this activity 16 years ago, I reasoned that if students actually took a test that appeared on the surface to be biased, the experience would stimulate their thinking and facilitate learning about important aspects of testing such as different types of test bias and different types of validity (e.g., face validity vs. predictive validity). Because of the controversial nature of intelligence tests, a fictional intelligence test seemed ideal. However, I needed one that did not take much class time, could be scored easily, contained engaging test items, and could be administered in both small and large classes.

After a lecture on problem solving, an introductory student showed me a copy of Omni Games: The Best Brains Teasers From Omni Magazine (Morris, 1983). It contained two "intelligence" tests (humorously labeled "College Entrance Exam 1" and "Advanced College Entrance Exam") that were ideal for the class activity. These tests involved solving rebus-type puzzles, commonly called "wacky wordies" (see Figures 1 and 2). I used these two tests with only minor changes.

Procedure

The activity requires minimal preparation. The two tests (Figures 1 and 2) and the answer sheets (given in the Appendix) can be copied for individual administration and feedback or made into transparencies for group use.

There are two parts to the activity. Because students should not know that there will be two tests, the tests should not be numbered as they are here for identification purposes. Instruct students to number 1 through 24 on a sheet of paper on which

![Figure 1. Twenty-four test items in Test 1.](image-url)
In this article we report the findings of a study (N = 18) that investigated the effectiveness of a teaching activity designed to enhance students’ understanding of the American Psychological Association (APA) manual and style (APA, 1994). Students read a poorly written paper and located as many style errors as possible. Students reported positive perceptions of the activity and results indicated students’ knowledge of style improved significantly from preactivity to postactivity quizzes. Furthermore, performance on the activity was significantly related to the application of APA style in an empirical report.

Although initially developed for use in psychological and anthropological journals, the format guidelines issued by the American Psychological Association (APA) have been adopted by a variety of publications and disciplines (Stahl, 1987). Students’ first training in APA style frequently occurs in research methods courses, although for many students mastering the appropriate editorial and writing guidelines specified in the Publication Manual of the American Psychological Association (APA, 1994) remains a struggle. It is an equally difficult task for instructors to design engaging activities to facilitate learning of APA style.

Several authors, however, have published techniques that may assist students in learning APA writing style. Stahl (1987) demonstrated that using checklists and uniform requirements across courses facilitated instruction in APA style. Ault (1991) designed an assignment in which students learned the appropriate organization of reports by taking scrambled paragraphs and placing them in the appropriate sections. Rosenthal, Soper, Coon, and Von Bergen (1999) developed a technique in which the instructor presented the first page of each student’s introduction on the overhead (without names) and edited each introduction with class input.

Although faculty members have developed a number of useful teaching activities addressing APA style, few activities were designed to encourage the appropriate use of the APA manual. One exception is a well-received training manual created by Gelfand and Walker (1990). This manual provides students with opportunities to test their knowledge of APA guidelines with multiple-choice quizzes and by employing these guidelines in exercises. The exercises in the work-
One week after the activity due date, students turned in their first APA-style research paper. The instructor removed student names from papers and evaluated papers using a 100-point scale. The evaluation of the research paper was based not only on APA style, but on content and writing proficiency. At this time, students took a 13-item postactivity quiz on APA style. The items on the preactivity quiz and postactivity quiz varied, although they addressed the same learning objectives and came from the same test bank (Leary, 1995).

Students also rated their agreement on a 9-point scale ranging from 1 (strongly disagree) to 9 (strongly agree) with 5 items that stated the APA style activity (a) helped them learn to use the manual, (b) increased their familiarity with the specific sections of the manual, (c) increased their knowledge of APA style, (d) would help them in the process of writing their own research report, and (e) was a fair assessment of their knowledge. Finally, students reported the number of hours it took them to complete the activity.

### Method

#### Participants

Students (13 women, 5 men) enrolled in a research methods class participated in the study. Six students were seniors, 10 were juniors, and 2 were sophomores. When asked about their prior experience with APA style, all students reported either having no experience or having used APA style “cheat sheets” when writing papers for other classes.

#### Procedure and Materials

To create the poorly written paper used for the activity, we adapted a paper written by an undergraduate research student (under the supervision of the authors). The study reported in the original paper concerned gender differences in attitudes toward controversial topics such as abortion rights and euthanasia. We altered the original paper so that it included a broad range of APA errors, such as errors with running heads, page numbers, verb tense, sexist language, heading levels, citations, references, and omissions in the results.

The course instructor gave a 40-min lecture addressing writing papers in APA style. Three days after the lecture, students completed a 13-item preactivity quiz. Five days after the lecture, students received a copy of the paper with the instructions to use the APA manual to locate as many errors as possible in the manuscript. Students had 5 days to complete the activity. Students reported errors by writing down the number of the page on which the error occurred, a description of the error, and the APA manual page number where the correct information could be found. Students earned 1 point for each type of error they found and reported in the appropriate manner. For example, if a student found an error in the font used for the running head that was repeated on each page of the manuscript, the student received only 1 point for documenting the error.

One week after the activity due date, students turned in their first APA-style research paper. The instructor removed student names from papers and evaluated papers using a 100-point scale. The evaluation of the research paper was based not only on APA style, but on content and writing proficiency. At this time, students took a 13-item postactivity quiz on APA style. The items on the preactivity quiz and postactivity quiz varied, although they addressed the same learning objectives and came from the same test bank (Leary, 1995).

Students also rated their agreement on a 9-point scale ranging from 1 (strongly disagree) to 9 (strongly agree) with 5 items that stated the APA style activity (a) helped them learn to use the manual, (b) increased their familiarity with the specific sections of the manual, (c) increased their knowledge of APA style, (d) would help them in the process of writing their own research report, and (e) was a fair assessment of their knowledge. Finally, students reported the number of hours it took them to complete the activity.

### Results

#### Student Perceptions of the Activity

As shown in Table 1, students expressed strong agreement that the activity helped them learn to use APA style and increased their familiarity with specific sections of the manual. Students stated that the activity increased their knowledge of APA style and would help them in the process of writing their reports. Twenty-eight percent of students reported that it took 3 to 4 hr to complete the activity, 33% reported that it took 5 to 6 hr, 22% reported that it took 7 to 8 hr, and 17% reported that the activity took more than 8 hr.

#### Activity and Quiz Performance

Students’ scores on the activity ranged from 21 to 66, and the mean score was 40.89 (SD = 13.22) of 70 possible errors to be located. The mean on the preactivity quiz was 7.33 (SD = 1.88) or 56% correct. The mean on the postactivity quiz was 9.72 (SD = 1.64) or 75% correct.

We performed a repeated measures ANOVA to test whether quiz scores were higher after participation in the activity and completion of the APA report. The within-subjects

| 1. To learn the APA Manual | 7.67 ± 1.19 |
| 2. To become familiar with sections of the APA Manual | 7.44 ± 1.25 |
| 3. To increase my knowledge of APA style | 8.44 ± 0.78 |
| 4. Writing of my own reports | 7.61 ± 1.29 |
| 5. The activity was a fair assessment of my knowledge | 5.89 ± 2.19 |

Note. Students’ perceptions rated on a scale ranging from 1 (strongly disagree) to 9 (strongly agree). APA = American Psychological Association.
variable was the timing of the quiz (preactivity and postactivity). The repeated measures ANOVA demonstrated a significant increase in quiz performance between the preactivity and postactivity quiz, $F(1, 17) = 24.16, p < .001, \eta^2 = .59$.

Application of APA Style in a Research Report

Student performance on the research report yielded a mean of 84.22 ($SD = 6.35$). Correlational analysis demonstrated that activity performance significantly related to the application of APA style, such that there was a significant positive correlation between activity scores and research report grade, $r(16) = .50, p = .02$.

Discussion

The findings of this study demonstrated that the activity was useful in teaching students about the APA manual and APA style. Students’ perceptions indicated that they found the activity to be helpful. In fact, students’ assessment of the activity was surprisingly positive given the number of hours the activity required.

Students’ performance on the activity was also related to the appropriate application of APA style in the research report. However, a number of other factors could influence report performance. For example, students’ final performance in a writing assignment will be influenced by both knowledge of APA style and individual differences in writing proficiency. We did not include an assessment of writing proficiency, so it cannot be ruled out as a spurious variable. Future studies could use preactivity and postactivity assessments of writing samples to investigate whether students’ writing performance improves after the activity.

There were increases in preactivity and postactivity quiz scores; however, these results should be interpreted with caution. Specifically, due to the study design we cannot determine whether increases in scores were due to the APA activity, the process of writing a research report, or a combination of both. Future studies could clarify the relative contributions of the activity and research report by systematically varying the timing of each in relation to the knowledge assessment. Other limitations of our study include the sample size and restriction of the sample to upper level psychology majors enrolled in research methods classes. Motivation regarding the course and activity may have been high.

Several alterations could be made to the procedure to allow instructors to use the activity in a variety of classroom settings. For example, instructors could use the activity with groups of students rather than as an individual assignment. Small groups of students could work as teams to locate errors with a group grade assigned for the errors located. One aspect of the activity that should not be altered is the way in which students record errors. The requirement of reporting manual page numbers is essential to familiarize students with the manual sections.

In conclusion, the activity described in this article was useful in encouraging student comprehension of the APA manual and style. Students reported through conversations with us that they felt much more prepared and less anxious in writing their own papers after participating in the activity. In this process of writing research reports, in many academic settings, students may refer to the manual to ensure that their writing reflects APA style. This activity may be of use to faculty who want to expand on this practice by requiring students to make concurrent comparisons between the manual and a manuscript.

References


Notes

1. This project was funded, in part, through a faculty development grant at Frostburg State University, Frostburg, MD.
2. The activity paper and the complete scoring template are available at the Web site for Elon University, Psychology Department: http://www.elon.edu/psychology/faceages.html.
3. Send correspondence to Gabie E. Smith, Department of Psychology, Campus Box 2163, Elon University, Elon, NC 27244; e-mail: gsmith@elon.edu.
In the description of the model, the teacher's terminology is critical. In one instance, a student used words such as “parallel,” “perpendicular,” and “flush against” to describe the relationship of two Duplo pieces. My students observed the learner pause to reflect on the difference in meaning in these concepts. They described this as a moment of disequilibrium in which the learner had to accommodate rather than efficiently assimilate information into existing schemata. As a result, the learner fell behind the teacher’s pace of instructions and became disoriented.

An example of assimilation leading to a faulty conclusion occurred on one occasion when one Duplo piece had an “eye-ball” imprinted onto the side of the block and the teacher described a “crescent piece” (similar to a half moon Duplo block) in her verbal directions. The learner's familiarity with Duplo led him to look for a crescent shaped piece that is often found in Duplo sets rather than a crescent shaped decal imprinted on the side of a square block.

Educational psychology students typically focus their observations on the learner rather than teacher, but as future educators, they should also attend to a teacher’s descriptions; instructions; use of advance organizers; and the techniques to aid assimilation, accommodation, and equilibration. This demonstration also relates to Gestalt theory’s central concept that the whole is greater than the sum of its parts. If learners do not have a complete visual or advance representation of the model, then their ability to gain information may be limited by descriptions focused on the component parts.

By not allowing the learner to ask any questions, this demonstration actually sets the stage for cognitive disequilibrium to occur. Learners can become frustrated that the teacher's directions are unclear or that vocabulary is uncommon. They can become bored by the slow pace of directions or agitated if instructions are too fast. When learners assimilate information incorrectly, the error affects the structure exponentially, as one piece of information is crucial for each component that follows.

Effectiveness

Students enjoy the demonstration and make several appropriate connections to the teaching environment as a whole; to Piaget’s classic theories; and to concepts such as advanced organizers, Gestalt theory, verbal dialogue, and pace of instruction. Students also experience a practice they should use in the future when they become teachers: converting theoretical concepts into real-world applications to promote intellectual engagement and motivation for learning. The real key is that students come to realize that instruction and learning are interactive activities involving both the teacher and learner.

Conclusions

This exercise in the discovery and application of Piaget’s cognitive theory encourages a reflective attitude in students who see and experience the disequilibrium that Piaget believed was necessary for learning to occur. Others (Burch, 1999; Grant, 1995; Hill, 2000) also recognized the importance of reflective practice as a means to understanding not only behavior but also cognitive development in the classroom. This self-reflective thought encourages a deeper self-understanding, which is an impetus to self-confidence, both of which are important traits for novice and seasoned teachers to understand and develop throughout their professional careers (Burch, 1999).

In conclusion, the classroom activity described in this article allows students to experience disequilibrium in a safe atmosphere and to explore Piaget’s concepts of schemata, assimilation, and accommodation in a practical setting. At the same time, it encourages the discussion of several additional important concepts that preservice and even current teachers should experience. As university instructors, we need to continually strive to develop classroom demonstrations and models that effectively assist our students to bridge the gap between theory and practice.

References


Notes

1. An earlier version of this article was presented at the Teaching Educational Psychology Symposium annual meeting of the American Educational Research Association, New Orleans, LA, April 2000.
2. Send correspondence to Joe D. Nichols, School of Education, Indiana University–Purdue University Fort Wayne, Fort Wayne, IN 46805; e-mail: nicholsj@ipfw.edu.

Using Student Scholarship to Develop Student Research and Writing Skills

Mark E. Ware
Amy S. Badura
Creighton University

Stephen F. Davis
Emporia State University

We illustrate the use of psychology student publications for teaching (a) principles in experimental methodology, (b) the use and in-
principles of experimental methodology, the use and interpretation of statistical tests, and the quality and format of American Psychological Association (APA) writing style. We present several exercises for each of these topical areas using published examples of student scholarship as teaching tools. We also identify published instances of errors in design, statistical use, and APA style. We contend that using students’ scholarly publications can promote and reinforce development of research and writing skills that combine active participation and critical thinking.

Several issues confronted participants at the 1991 National Conference for Enhancing the Quality of Undergraduate Education in Psychology (McGovern, 1993). Two of the major themes for improving education emphasized increasing students’ active participation in learning (Mathie et al., 1993) and critical thinking skills (Halpern et al., 1993). Since then, several authors have discussed strategies for improving active learning and critical thinking (Halonen, 1995; Henderson, 1995; Hubbard & Ritchie, 1995; Perry, Huss, McAuliff, & Galas, 1996; Seegmiller, 1995; Wade, 1995). Yet another strategy combining active participation and critical thinking consists of in-class exercises using published student scholarship.

Several educators have reported on the value of professional literature as a teaching device. Suter and Frank (1986) used published literature to illustrate core concepts to students. Hubbard and Ritchie (1995) stimulated critical thinking through the assessment of scholarly work, and Carkenord (1994) motivated students through the reading of professional literature. At a more fundamental level, Pennington (1992) had students read an abstract or parts of a method section to understand concepts such as independent and dependent variables.

Although the use of published articles as teaching tools in the classroom is not new, this article’s emphasis on the value of student publications is relatively uncommon. One of the few examples presenting this technique is a research methods text (Smith & Davis, 2001). A review of student journals indicates a high degree of similarity between student and professional publications in terms of variety of topics, variables, and statistical analyses. However, student research reports tend to be shorter, more simplistic, and easier to comprehend than traditional professional articles. All these benefits allow educators to illustrate core concepts directly and efficiently. In addition to providing stimuli for engaging class exercises, published student articles also meet undergraduates’ desires for models of excellent student work to guide them in developing new skills. Therefore, student literature meets instructors’ needs to engage students in active learning and critical thinking and, at the same time, meets students’ needs to have instructors provide clear examples of outstanding undergraduate performance.

Student articles can be effective teaching aids for a variety of research and writing skills. Published student articles are readily available as adjuncts to traditional reading assignments. Educators have established several journals devoted to the publication of undergraduate student research, such as the Journal of Psychological Inquiry (JPI), Psi Chi Journal of Undergraduate Research (PCJUR), and The Journal of Psychology and the Behavioral Sciences (JPBS). In this article, we illustrate how instructors can use published student scholarship for teaching principles of experimental methodology, the use and interpretation of statistical tests, and the quality and format of American Psychological Association (APA) writing style.

### Exercises

Using articles appearing in the student journals listed previously, we developed exercises for each of three types of courses: research methods; statistics; and topical, content-based psychology courses. Instructors can use some exercises, particularly those involving writing assignments, for more than one course. Students can complete each of the exercises in this article within 10 to 25 min, making them ideal for use throughout the semester.

### Research Methods or Experimental Psychology

An instructor can conduct this exercise when students are learning the fundamentals of research including independent, dependent, and extraneous variables. The instructor can assign an article as required reading one class period before conducting the exercise or require students to read the article during class; students need read only the introduction and method sections. Independently or in small groups, students can identify manipulated as well as measured variables. A group discussion about the identified variables would follow. We used Sheets (1999) to illustrate identification of independent variables (extraverted or introverted behavior; sex of participant), dependent variables (several Likert-type scales), control procedures (identical stimuli across conditions), and experimental design (2 × 2 ANOVA).

### Statistics

For statistics, we found Bleeker, Evans, Fisher, and Miller’s (1998) article particularly helpful. We had students (a) determine one of the primary statistical tests (3 × 2 ANOVA), (b) determine the appropriate number of degrees of freedom, and (c) compare the calculated statistic with the appropriate tabled statistic. With the information available in Bleeker et al. (1998), students can confirm the accuracy of the reported ANOVA degrees of freedom and the decision to reject the null hypothesis.

Students may also discover that reviewers and editors failed to find mistakes in reported results. Verbeck (1996) contained one error in the degrees of freedom because one student’s data were excluded. Recognizing such errors can inform students about the importance of developing their own critical thinking and copyediting skills as well as educating them about degrees of freedom.

Instructors can also use Bleeker et al. (1998) to help understand another difficult concept: interaction. Instructors can ask students to identify, evaluate, and explain the main effects and interaction findings for the self-esteem two-factor ANOVA. Neither main effect was significant; however, the interaction between Gender × Group Participation was significant. Examining Figure 1 (Bleeker et al., 1998, p. 36), which showed the mean self-esteem scores, may help students understand the meaning of an interaction, as can an
examination of the discussion section that presented the statistical findings concretely.

Topical Courses with Writing Assignments

When students complain about difficulty in writing psychology manuscripts, they usually complain about writing fundamentals and about APA format. Other educators have described strategies for improving the quality of student’s literature reviews (Froese, Gantz, & Henry, 1998) and experimental reports (Ault, 1991; Dunn, 1996, 1999; Peden, 1994; Sternberg, 2000). Web sites list common violations in writing style among undergraduates (e.g., http://puffin.creighton.edu/psy/journal/frequerr.html) and offer self-tests that incorporate many of the more common errors in style and language (e.g., http://www.lemoyne.edu/OTRP/otrpresources/otrpsciwriting.html). Dunn et al. (2001) offered a checklist of common formatting errors for manuscripts submitted to the PCJUR.

Expression of ideas. This exercise focuses on avoiding students’ common grammatical errors. Instructors can direct students to examine published articles for examples of proper use of difficult grammatical rules. For example, students often have particular difficulty grasping the difference between active and passive voice. APA style experts generally prefer active voice. Peluso (2000) appropriately used active voice throughout the introduction, clearly identifying the actors in the sentence as the researchers. Students can also examine Peluso’s method section to identify examples of acceptable use of passive voice when the focus of the sentence is “on the object or recipient of the action rather than on the actor” (APA, 2001, pp. 41–42).

Mechanics of format: Title and abstract. Two simple in-class exercises involve asking students to check for APA format by counting the number of words in an article’s title and its abstract. According to the Publication Manual (APA, 2001), title length should be 10 to 12 words, and abstract length of a review article should be from 75 to 100 words. West and Berning’s (1999) JPBS article closely approximated both criteria.

Mechanics of format: References. In this exercise, instructors can direct students to count the number of times each item in the references is cited in the article. Although the exercise has nothing to do with the frequency with which authors cite items, students can determine whether the article conforms to the APA requirement that all references cited in the text should appear in the reference list and vice versa. West and Berning (1999) cited all reference list items in the article.

In another, more advanced APA-format exercise, instructors can give groups of students an article and direct them to produce the manuscript version for various sections, such as the title or the references. Many students are surprised at the differences between the published and manuscript versions.

With respect to referencing, students seem to have great difficulty mastering the proper use of “et al.” and page numbers when referencing in the text. Students can examine the West and Berning (1999) article and discover on the first page an example of a proper use of “et al.” after all the authors have been cited once. Students can also note that page numbers were used with direct quotes.

On the second page, students will find a correct citation of “Zametkin et al. (1990)” denoting the first reference to that work in the article. Instructors can use this example to illustrate the proper use of et al. when a published work has six or more authors. However, the six or more author rule was violated on the third page of the article, in which West and Berning (1999) incorrectly cited “(Rounsaville, Anton, Carroll, Budde, Prusoff et al., 1991)” as well as omitted the comma that belongs between the last author’s name and et al. This erroneous citation demonstrates the difficult nature of mastering complicated citation rules and instructs students to identify and avoid format pitfalls before they complete major writing assignments. Finally, this example also illustrates that editors, reviewers, and copyeditors can overlook such errors.

Conclusions

The exercises and specific student reports outlined in this article are only a sample of the possible exercises and published student work available to educators. Instructors can purchase individual journals for classroom use, obtain copyright permission very inexpensively for use of published articles from JPBS, PCJUR, or encourage students to view past issues of JPBS online. Further information about student research publications is readily available at Web sites for the three journals discussed in this article (JPBS, http://puffin.creighton.edu/psy/journal/JPIhome.html; PCJUR, http://www.mercyhurst.edu/UPD/UPDdescriptions.htm#PsiChi; JPBS, http://alpha.fdu.edu/psychweb/JPBS.htm). Additionally, a Society for the Teaching of Psychology Web site (http://www.lemoyne.edu/OTRP/otrpresources/otrpteachingformat.html) contains information about those and other publication opportunities for undergraduate students.

As a final encouragement for educators to include some of the exercises described in their research methods, statistics, or content-based courses, we summarize some of our students’ observations about these exercises. Students reported that they learned how to apply experimental concepts and statistics to research, format statistical analyses in written text, and improve their writing style. In summary our experience suggests that students can learn concepts, principles, and writing style from student-generated research. Future research should formally and systematically investigate the pedagogical opportunities that these and similar exercises offer for developing student research and writing skills.

References


Notes

1. We thank Randolph A. Smith and four anonymous reviewers for their original ideas and important contributions to this article.

2. Send correspondence to Mark E. Ware, Department of Psychology, Creighton University, Omaha, NE 68178; e-mail: meware@creighton.edu.

**Coverage of Industrial/Organizational Psychology in Introductory Psychology Textbooks: An Update**

Douglas C. Maynard
Karen L. Geberth
Todd A. Joseph
State University of New York at New Paltz

Carlson and Millard (1984) found that Introductory Psychology textbooks provided limited if any coverage of industrial/organizational (I/O) psychology. We examined current Introductory Psychology textbooks (N = 54) for the presence of a section, appendix, or chapter on I/O psychology. We also coded textbooks for the number of pages containing I/O content. Results were similar to those found previously; only about one fourth of textbooks contained an overview of the field in some form. Full-length textbooks were more likely than brief versions to contain an I/O section, appendix, or chapter. On average, less than 2% of the total number of textbook pages contained work-related concepts or examples; this percentage was similar for full-length and brief textbooks. We justify the importance of increased coverage of I/O psychology in future text editions.

Student exposure to industrial/organizational (I/O) psychology in Introductory Psychology courses would be beneficial for several reasons. First, because more psychology departments are offering I/O courses (Perlman & McCann, 1999), early exposure to the field would help students make more informed decisions about taking an I/O course. Second, instructors can use I/O psychology examples to demonstrate how psychologists apply core psychological concepts (e.g., motivation) to real-world problems. Third, psychology undergraduates are more likely to find employment in business and management than any other occupational area (“A Look at Recent Baccalaureates in Psychology,” 2000). Fourth, the general public lacks awareness of the field of I/O psychology (Gasser et al., 1998); student exposure to I/O in a course as popular as Introductory Psychology would help to ameliorate this problem.

Textbook coverage of I/O psychology, which is outside of the expertise of most instructors, may determine the likelihood that these instructors discuss the field during class time. Carlson and Millard (1984) conducted a content analysis of...
Implementing and Evaluating a Writing Course for Psychology Majors

Perilou Goddard
Northern Kentucky University

In this article, I describe Writing in Psychology, a semester-length 3-credit elective course designed to improve students’ writing skills, familiarize them with psychology’s writing conventions, and teach them American Psychological Association (APA) style. Students produced a case report, a report of an empirical study, a conference abstract, and a literature review. An attitude inventory and tests over grammar and APA style revealed significant precourse versus postcourse improvement, providing evidence that the course can be a valuable addition to the undergraduate curriculum.

Many students struggle to learn to write well in psychology’s technical style. Even students who regularly earn As in English classes often are shocked to receive critical feedback on their psychology papers (Nadelman, 1990). However, acquiring writing skills relevant to psychology is a crucial aspect of becoming socialized into the discipline (Madigan, Johnson, & Linton, 1995). As McGovern and Hogshad (1990) observed, learning to write reflects students’ ongoing cognitive development in their psychology courses and in their college careers more generally.

Psychology faculty increasingly face pressure to improve their students’ writing skills. Since its inception in the mid-1970s, the writing-across-the-curriculum movement has insisted that teaching writing is not the exclusive purview of English Department faculty (Fulwiler & Young, 1990; Rickabaugh, 1993). Many psychology faculty have responded to this movement, as well as to their perceptions of students’ writing deficiencies, by increasing the writing demands and range of writing assignments in their courses (e.g., Connor-Greene, 2000; Dunn, 2000; Henderson, 2000; Nodine, 1990a; Norcross, Slotterback, & Krebs, 2001). Many of these assignments use writing to help students learn psychology; other assignments focus on helping students learn to write like psychologists (Nodine, 1990b). The latter goal is pursued in most research methods courses, which continue to emphasize learning American Psychological Association (APA) style. Nevertheless, few faculty appear to have answered Calhoun and Selby’s (1979) call to teach a course specifically focused on writing in psychology.

In 1999, I initiated Writing in Psychology, a semester-length three-credit elective course focused on improving students’ writing in general as well as teaching them APA style and psychology’s writing conventions. My course has several features in common with Rileigh’s (1998) Communication Skills in Psychology course. For example, both courses cover use of library resources and acquisition of skills in topic selection, grammar, organization, and draft revision. However, my course focuses on teaching students to produce full-scale versions of the types of writing assignments required most often in undergraduate psychology courses: case reports, reports of empirical studies, conference abstracts, and literature reviews. Like Rileigh, I evaluated the course by assessing the change in students’ understanding of APA style. I also evaluated students’ improvement in basic grammatical conventions and assessed their change in attitudes toward course-relevant objectives.

Method

Participants

Eligible participants included all students enrolled in Writing in Psychology during two consecutive semesters (N = 29; 17 in Fall 1999, 12 in Spring 2000) at a midsize metropolitan university in the Midwest. Two students (one in each semester) were absent during the postcourse assessment; their data are not included in the analyses. All students were psychology majors or intended to declare psychology soon after completing the course. Prerequisites included successful completion of College Writing (the freshman composition course), Introduction to Psychology, and at least one additional psychology course. Students represented all classifications (3 freshmen, 13 sophomores, 4 juniors, and 7 seniors). With the participants’ consent, I obtained their composite American College Testing (ACT) scores from university records. Their median ACT score was 20 (the same as the university’s median); however, the scores ranged from 10 (approximately the 1st percentile nationally) to 30 (97th percentile nationally). The median number of psychology courses taken prior to the writing course was 6 (range = 2 to 16). Students could take the writing course before, during, or after taking the required block of research methods courses; 18 took the course prior to taking Research Methods and Tools, 3 took the course concurrently with the methodology courses, and 6 took the course after having completed the methodology block. Thus, the participant sample was heterogeneous with regard to preparation for college-level work, prior exposure to psychology coursework in general, and prior exposure to research methodology specifically.

Course Content

Writing in Psychology fulfilled university requirements for an advanced composition course and also counted as an elective course in psychology. Teaching methods included short...
exercises on grammar and APA style, editing of drafts by peers, and revision of drafts following instructor comments. Required textbooks included the *Publication Manual of the American Psychological Association* (APA, 1994); Parrott’s (1999) *How to Write Psychology Papers*; and *Rules of Thumb: A Guide for Writers*, a grammar and usage text by Silverman, Hughes, and Wienbroer (1999b). Primary assignments included a case report, a report of an empirical study, a conference abstract, and a literature review. Students completed all assignments except the conference abstract in a progressive format (Hemenover, Caster, & Mizumoto, 1999). That is, students accomplished each project in stages, with multiple opportunities to receive feedback on their progress and revise their writing as needed. Furthermore, in contrast to the recommendations of some authors (e.g., Price, 1990; Willingham, 1990), I took an active stance regarding feedback on the mechanics of students’ writing. I indicated (but generally did not correct) mechanical errors in every assignment submitted to me and gave frequent brief lessons on specific types of grammatical problems.

**Case report.** The main purpose of the case report, the first major assignment students completed, was to give students an interesting opportunity to learn several conventions of professional communication that did not require APA style. The assignment also gave me the chance to assess students’ baseline writing skills, establish norms regarding the appropriate tone of the writing, and introduce the process of peer editing. For the case report, I told students to assume the role of a clinical psychology intern nearing the end of internship training who had conducted four sessions of cognitive behavioral therapy with a specific client (“Lisa”). I instructed students to prepare a report for the intern who would take over as Lisa’s therapist.

I provided brief background information from the instructional materials included with the cognitive behavioral therapy videotape from the APA *Psychotherapy Videotape Series* (APA, n.d.). I then played the tape, asking students to imagine themselves in the role of the therapist. I instructed students to include in their case reports background information about Lisa, a description of her presenting problem, a summary of previous therapy sessions, and a summary of the most recent session (shown on the video). I also told them to make recommendations for continued therapy. My case report guidelines required students to (a) strive for a professional tone, (b) focus on describing the client’s behavior and self-reports (rather than making unsupported inferences), (c) avoid labeling the client, and (d) never use pejorative terms to describe the client.

I assigned students to small groups of three or four peers for editing and, as with all graded assignments in the course, students first prepared drafts for their colleagues. After addressing their peers’ comments, students submitted drafts to me for comments; I encouraged students to respond to my comments prior to submitting their final drafts for a grade.

**Report of an empirical study.** Because my department requires all psychology majors to take a research methods sequence, my focus in the writing course was not on teaching students to design and conduct studies. Instead, my goal for this assignment was to teach students the elements of an empirical report. In particular, I stressed how APA style contributes to the clear communication of empirical research and how the organization of an APA-style empirical report reflects the ideal research process (locating the research question in the context of past empirical work, systematically collecting and analyzing new data, and relating the findings to current theories; Madigan et al., 1995). To these ends, I designed, conducted, and analyzed a simple experiment applying the mere exposure effect (Zajonc, 1968) to a consumer psychology context. Introductory psychology students rated the ease of pronunciation of 12 fictitious candy bar names; 1 name was repeated four times during this task. Later, participants rated how much they would like each of the candy bars. As predicted, the more frequently repeated name yielded significantly higher ratings than the other names. I explained the study to the writing students and provided them with a set of articles they could use to construct a logical introduction. Each student then constructed a title page, introduction, and references for the paper, based on guidelines I provided.

I discussed and modeled all elements of the paper in detail, but I paid particular attention to the structure of the introduction, starting with broad statements and references to literature, narrowing the focus to the current study, and ending with the specific hypothesis. I sought to show my students that the persuasiveness of the introduction stems from the logical flow of one’s arguments, rather than from attention-getting prose (Sternberg, 1993).

After students revised their drafts of the title page, introduction, and references, they drafted the method, results, discussion, and abstract of the paper and submitted them for both peer and instructor comments. Finally, students submitted a complete report of the entire empirical study for a grade. Throughout our work on this assignment, I attempted to help students understand the elements of APA style necessary to prepare the report accurately. However, my main interest was in showing students how APA style enhances the precision of written communication; I explicitly advised them not to be concerned about memorizing the details of APA style.

**Conference abstract.** Because Dunn (1994) recommended giving students assignments that simulate professional experiences, and because many of our students submit abstracts to local, regional, or national conferences during their undergraduate careers, I had students work in groups to convert their individual empirical reports into collaboratively produced conference abstracts. Students grouped themselves into clusters of two to four members. They reviewed several examples of conference papers (both the brief abstracts published in conference programs and the longer conference abstracts submitted for review) and divided the writing tasks among themselves. Given that the previously graded empirical reports served as the basis for the conference abstracts, students did not submit rough drafts of their conference abstracts to me for review. Instead, each group submitted its own final conference abstract for a grade.

**Literature review.** Unlike the other course assignments, the literature review allowed students to choose their own topic. The assignment culminated in a paper analyzing at least
seven journal articles detailing original research; I permitted inclusion of other reference materials such as published literature reviews only if students met the requirement for the minimum number of original research articles. The literature review was the final graded assignment of the semester. However, following the recommendations of many authors (e.g., Boice, 1990; McGovern & Hogheaud, 1990; Nadelman, 1990), work on the review proceeded in stages throughout the semester.

First, following the recommendations of Boice (1990), Dunn (1994), and Poe (1990), I had students engage in a freewriting exercise in class to brainstorm possible topics for their reviews. Students then collaborated with their peers and me to narrow their topic to one that would likely yield a viable paper.

Next, the students met with an instructional librarian to learn to use the APA PsycINFO database to identify relevant journal articles. Students submitted an annotated version of the PsycINFO search to me for review. After obtaining the articles they planned to use in their review, students analyzed the quality of the empirical studies using guidelines I had compiled from recommendations of Maher (1978) and Meltzoff (1998).

Following Sternberg’s (1993) recommendations, I coached students on how to take notes on what they read and how to construct an outline for their paper. Students submitted their outlines and the articles they planned to use to me for comment. Finally, students drafted the literature review and submitted it first to their peers, and then to me, for recommendations before turning in the final paper for a grade.

**Evaluation Methods**

I explained to students that the course was new to me, and I wished to collect several types of data to evaluate the course’s effectiveness; however, I did not require students to participate. All students gave their informed consent to participate in the evaluation activities, which included precourse and postcourse tests on grammar and APA style and a 14-item Likert-type inventory assessing attitudes toward writing-relevant behaviors. Students’ participation in the evaluation activities did not affect their grades.

I constructed the grammar test from items in the practice manual (Silverman, Hughes, & Wienbroer, 1999a) accompanying the grammar textbook used in the course. I employed the same items for pretesting and posttesting. I reviewed the answers to the pretest briefly during the first week of class; however, students did not retain their copies of the pretest and I made no further reference to the specific items during the remainder of the semester.

Two different mastery tests from Gelfand and Walker’s (1990; revised in 1994) APA style guidebooks comprised the APA pretests and posttests. I did not permit students to consult the *Publication Manual* (APA, 1994) during the tests. After each test, I allowed students to review their responses with the answer key, but few opted to do so, perhaps because their performance on the tests did not affect their grades.

**Results**

To determine whether I could combine the responses to the attitude inventory into a single score, I reverse scored the negatively worded items (1 and 7) and computed Cronbach’s alpha coefficients on the precourse and postcourse responses. Because the internal consistency of the attitude inventory was relatively low (Cronbach’s $\alpha = .59$ precourse, .57 postcourse), I examined the 14 items individually. Paired-sample $t$ tests revealed significant ($p \leq .05$, two-tailed) differences in the predicted direction on 9 of the 14 items, with a nonsignificant trend on one additional item (see Table 1).

<table>
<thead>
<tr>
<th>Inventory Item</th>
<th>Precourse $M$</th>
<th>Postcourse $M$</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. I have confidence in my ability to write a report of an empirical study in psychology.</td>
<td>2.74</td>
<td>4.37</td>
<td>$–6.80$</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>3. I know how to use PsycINFO well.</td>
<td>2.00</td>
<td>4.52</td>
<td>$–9.02$</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>4. I have confidence in my ability to use the Internet to find useful information for research papers in psychology.</td>
<td>3.52</td>
<td>3.96</td>
<td>$–2.06$</td>
<td>.05</td>
</tr>
<tr>
<td>5. I know how to write a high-quality literature review in psychology.</td>
<td>2.04</td>
<td>4.15</td>
<td>$–8.91$</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>6. I know how to evaluate the quality of information I find on the Internet.</td>
<td>3.15</td>
<td>3.77</td>
<td>$–2.31$</td>
<td>.029</td>
</tr>
<tr>
<td>7. My writing is grammatically correct most of the time.</td>
<td>3.41</td>
<td>3.37</td>
<td>0.19</td>
<td>ns</td>
</tr>
<tr>
<td>8. Writing detailed outlines helps me write better papers.</td>
<td>3.22</td>
<td>3.26</td>
<td>$–0.21$</td>
<td>ns</td>
</tr>
<tr>
<td>10. I have a clear understanding of APA style.</td>
<td>2.00</td>
<td>3.70</td>
<td>$–6.54$</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>11. Writing more than one rough draft makes my papers better.</td>
<td>3.81</td>
<td>4.63</td>
<td>$–4.23$</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>12. My papers are better if I have others read my drafts and make suggestions for improvement.</td>
<td>4.07</td>
<td>4.78</td>
<td>$–4.21$</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>13. I know how to put information I’ve read in journal articles into my own words.</td>
<td>3.59</td>
<td>3.96</td>
<td>$–1.73$</td>
<td>.096</td>
</tr>
<tr>
<td>14. I know how to pick out the most important points when I read articles in psychology journals.</td>
<td>3.33</td>
<td>4.22</td>
<td>$–4.25$</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>1. I tend to put off writing assignments until the last minute. a</td>
<td>3.15</td>
<td>2.93</td>
<td>1.10</td>
<td>ns</td>
</tr>
<tr>
<td>7. I have difficulty understanding psychology journal articles. a</td>
<td>2.89</td>
<td>2.56</td>
<td>1.18</td>
<td>ns</td>
</tr>
</tbody>
</table>

Note: All items were rated on 5-point Likert scales ranging from 1 (strongly disagree) to 5 (strongly agree). All comparisons were two-tailed, paired-sample $t$ tests ($df = 26$, except Items 5 and 6: $df = 25$).

For ease of interpretation, the two negatively-worded items are grouped together here, but they appeared in numerical order when students completed the inventory.
The maximum possible number of correct answers on the grammar test was 33. The precourse mean was 22.93 (SD = 5.25); the postcourse mean improved to 26.19 (SD = 4.06). A two-tailed paired-sample t test indicated that the difference was significant, \( t(26) = -4.60, p < .001 \).

The maximum possible number of correct answers on the APA style tests was 40. The precourse mean score was 20.74 (SD = 3.80); the postcourse mean score improved to 23.85 (SD = 4.23). Again, a two-tailed paired-sample t test revealed a significant difference, \( t(26) = -3.68, p = .001 \).

The extreme heterogeneity of the sample with regard to ACT scores raised the question of whether all students benefited from the course. To address this question, I computed change scores (postcourse minus precourse) for both the grammar test and the APA style test. Not surprisingly, composite ACT scores correlated significantly with precourse and postcourse scores on both tests (\( r \) ranged from .53 to .65, \( ps \) ranged from .012 to .001). However, ACT scores were not significantly correlated with change on the grammar test, \( r(23) = -.10, p = .66 \), or change on the APA style test, \( r(23) = -.06, p = .78 \). Thus, it appears that students at all levels of ability improved their grammatical and APA style skills, at least insofar as such skills were reflected in their scores.

Discussion

This evaluation provides evidence that a psychology-specific writing course can be a valuable addition to the undergraduate psychology curriculum. As indicated by the writing attitudes inventory, students became significantly more confident in their ability to write empirical reports and literature reviews. They also became more sure of their skills in using APA style, conducting PsycINFO searches, and evaluating journal articles and information found on the Internet. Students’ attitudes toward adaptive writing behaviors, such as writing multiple drafts and having others read and critique their drafts, also became significantly more positive. Additional evidence for the effectiveness of the course comes from the statistically significant improvements on the grammar and APA style assessments and the observation that improvement on these tests was not correlated with students’ ACT scores. Although the impact of the course on samples of student writing remains to be demonstrated, the statistically significant improvements on the grammar and APA tests are particularly noteworthy, given that students’ grades were not contingent on their performance on these tests.

One remaining challenge regarding Writing in Psychology concerns the ideal placement of the course in the psychology curriculum. Students who had already taken the research methods sequence appeared to find the course easier than those who had not taken Research Methods and Tools, but students who took the writing course first may in turn find the research methods courses easier. Unfortunately, additional data analyses failed to clarify the issue. I found no statistically significant correlations between the number of previous psychology courses students had taken and any measure of course outcome. Moreover, in a series of t tests comparing the outcome variables of students who had already taken or were concurrently enrolled in Research Methods and Tools versus those who had not yet taken Research Methods and Tools, I found no statistically significant differences in the final course grades or grammar and APA test scores.

Another challenge is determining whether the course should be required of all psychology majors. Given that our department currently has approximately 350 majors, the practical reality is that this course cannot be required unless the department is willing to commit significant resources to it and other faculty members want to teach it. One possible compromise might be to recommend the course especially strongly to students who aspire to attend graduate school.

It is also apparent from the postcourse grammar and APA test scores that my students still had much to learn. Students’ understanding that their scores on those tests would not affect their course grades may have attenuated their test performance; the fact that I did not encourage my students to memorize APA rules but administered the precourse and postcourse APA tests in a closed-book format may also help explain why postcourse scores remained low. Nevertheless, it remains clear that my students’ and my efforts (including approximately 12 hr of grading per student) were not sufficient to turn most of them into accomplished writers.

A further issue relevant to this course is the extent to which instructors should focus on teaching the mechanics of writing. Willingham (1990) argued that mechanics are strongly emphasized in high school, students already know that mechanics matter, and the best way to get students to write with technical competence is to insist that they do so, requiring students to acquire such competence on their own if they do not already possess it. Although I agree with Willingham that correcting students’ errors is not the best way to help them learn, I am not convinced that my students were fully aware of their technical shortcomings, nor am I convinced that my students were sufficiently motivated or able to achieve technical competence on their own. Consistent with this argument, it is interesting to note that the only attitude item with means in the opposite direction from those predicted was Item 8 (“My writing is grammatically correct most of the time”). Although the difference in means was not statistically significant, the pattern of means on this item is not surprising; the course appears to “shake up” some students’ unwarranted confidence in their command of writing mechanics.

I found Writing in Psychology to be a challenging course to teach, in part because students complained frequently that APA style is rigid and boring. However, by emphasizing the format and organization dictated by APA style, rather than requiring students to memorize the details, I attempted to help students see that this style communicates the logic of the scientific process (Madigan et al., 1995). I focused on showing students that, like other sciences, psychology uses language in an efficient, utilitarian way and that the organization of psychological reports mirrors the ideal research process.
References


Notes

1. An earlier version of this article was presented at the eighth annual American Psychological Society Institute on the Teaching of Psychology, Toronto, Ontario, Canada, June 2001.

2. I thank Cyndi McDaniel, James H. Thomas, Frank Kardes, and three anonymous reviewers for their helpful comments on earlier drafts of this article.

3. Send correspondence to Perilou Goddard, Department of Psychology, Northern Kentucky University, Highland Heights, KY 41099; e-mail: goddard@nkku.edu.

Notes

1. Development of the archive and writing of this article were greatly facilitated through the support of a Middle Tennessee State University Faculty Research and Creative Activity Committee Grant awarded to Tom Brinthaupt and John Pennington.
2. We thank Doug Cothern, Dennis Kramer, and Will Langston for their comments on an earlier version of the manuscript.
3. Readers can view the archive site at the following address: http://mtsu32.mtsu.edu:11311/archive/psych_data_archive.htm.
4. Send correspondence to Tom Brinthaupt, P.O. Box X034, Department of Psychology, Middle Tennessee State University, Murfreesboro, TN 37132; e-mail: tbrintha@mtsu.edu.

'Tis Better to Give Than to Receive: An Undergraduate Peer Review Project

Theresa L. White and Brenda J. Kirby
Le Moyne College

Forty-six undergraduates studying research methods had the opportunity to learn about the publication process by mimicking submission to a journal. Students then acted as blind reviewers for their peers’ papers and received blind reviews of their own. Based on the reviews, students edited and resubmitted their papers. They indicated that they gained more knowledge through writing the reviews than by receiving them.

Research methods students learn that the scientific method forms the basis for viewing psychology as a science. One component of the scientific method is communication of findings to the professional community, which typically involves publication in a peer-reviewed journal. With few exceptions, the peer review process improves manuscripts submitted for publication. In this manner, peer review both protects and enhances science through constructive criticism and collective knowledge.

This article describes a peer review exercise for use with undergraduate students in an introductory research methods course. The goals of this exercise are to allow students to develop writing abilities, learn about the publication process, and to enhance their learning of research methodology. Students completed a group project and then had a writing opportunity similar to a professional psychologist in the submission of papers for blind peer reviews and revision of their manuscripts. We derived this project from a published peer review activity with very different goals in which graduate students in content courses submitted independently for blind peer reviews and revision of their manuscripts. We derived this project from a published peer review activity with very different goals in which graduate students in content courses submitted independently for blind peer reviews and revision of their manuscripts. We derived this project from a published peer review activity with very different goals in which graduate students in content courses submitted independently for blind peer reviews and revision of their manuscripts.

Giving constructive criticism, the foundation of the peer review process, is a difficult skill to develop. A reviewer can be overly critical and offensive rather than helpful. In an effort to foster reviews that would be professional, impersonal, and offer honest, direct advice, students read about writing fair reviews (e.g., Sternberg, 2002a, 2002b) before beginning the peer review project.

Method

Participants

Forty-six research methods students enrolled in four sections (two sections taught by each author) participated as partial fulfillment of course requirements.

Materials

A seven-item checklist questionnaire assessed student reviewer opinions of a peer article (see Table 1). An additional questionnaire assessed student opinions of the peer review project through Likert ratings (see Table 2).

Procedure

The class completed two studies (one observational, one experimental), with students preparing manuscripts for each using the Publication Manual of the American Psychological Association (American Psychological Association [APA], 2001) as though for submission to a journal.

To ensure blind review, the instructors removed the cover pages from two of the three submitted copies of each manuscript and labeled them with preassigned identification numbers that varied between the two class projects to protect anonymity. Each student completed a checklist-style questionnaire (see Table 1) along with a peer review (maximum two pages plus cover sheet) for two randomly assigned manuscripts. Students also wrote comments directly on the reviewed papers. The instructors evaluated each peer review for effectiveness of content, corrected incorrect information, and returned to the reviewers the cover pages with comments and grades. Student authors received three reviews from their initial submission: two from student peer reviewers and one from the course instructor. Students revised and resubmitted their manuscript within a week for an additional grade. At semester’s end, students completed anonymous ratings of the peer review process (see questions in Table 2).

Results

Two student reviewers completed the checklist evaluation for each submitted manuscript to allow calculation of a percentage agreement intrarater reliability measure. Agreement varied widely by question, ranging from 35.6% to 84.8%, indicating that students differed in their abilities to assess specific aspects of the publication process. The highest agreement level occurred when students assessed whether the manuscript author understood the experiment, and the lowest agreement was in evaluating whether there were errors in APA style (see Table 1). Pearson’s correlation coefficients revealed no significant correlation in overall agreements between the two student reviewers on either paper: observation paper, r(44) = .11, p = .47; experimental paper, r(43) = .05, p = .76.
The evaluations of the accuracy, formatting, and writing quality on a 30-point scale of each student’s peer reviews give a rough indication of students’ abilities to assess the work of their peers. This ability marginally improved from the first (M = 26.94, SD = 2.20) to the second (M = 27.67, SD = 1.83) lab review, t(45) = 2.00, p = .05.

Forty students evaluated the submission of their papers for peer review. Mean responses to each question appear in Table 2, as do single-sample t tests for each question that compare the given responses to neutral responses (5 on a 0- to 10-point scale). It is noteworthy that students were neutral only in their responses to questions dealing with the helpfulness of receiving information from their fellow students. Most students indicated that the peer review project helped them understand the publication process and that they would not rather just submit a single paper instead of participating in the exercise. The ratings showed that students believed that reviewing the papers of their peers was more helpful than receiving reviews from their peers, paired t(39) = 4.22, p < .001. Students also indicated that seeing and reviewing the papers of others improved their papers more than receiving reviews from their peers, paired t(39) = 3.55, p < .001.

**Discussion**

Most students indicated that the peer review process helped them understand the publication process and that they preferred it to simply submitting a single paper. Thus, the peer review project provided a writing experience for undergraduate students that taught the advantages of the peer review aspects of publication including the benefit of revising and resubmitting a paper.

The students indicated that reviewing improved their papers more than receiving peer comments, which may reflect the inconsistency of information conveyed by peers. At least initially, students had a poor understanding of a professional APA manuscript, and increasing this understanding was one of the goals of this exercise. The uneven learning across students may account for the low interrater reliability level reported in this study, although it is not vastly inconsistent with published accounts of interrater reliabilities associated with professional reviewers (Marsh & Ball, 1989). A lack of reliance on information contained in peer reviews may also reflect students’ keen awareness that the final evaluation would come from the instructor. Just as professional psychologists afford more weight to a journal editor’s suggestions than those of individual reviewers, students may attend more to information that they know with certainty will improve their grades.

Although Haaga (1993) described a similar peer review activity for graduate students as a way to promote publication and development of professional identity, the literature rarely contains reports of such exercises with undergraduate students (Dunn, 1996). Nevertheless, peer review activities are possible with students as young as the high school level (Gift & Krasny, 2003), and online applications are available to assist in developing peer review skills (Robinson, 2001).

Despite the increased work load for the instructor of roughly 5 hr per project, this research indicates that undergraduates benefit from peer review. They learn through teaching others about publication and prefer the “revise and resubmit” nature of the process to writing one paper. By mimicking the journal publication process, undergraduates learn about APA style, receive additional exposure to writing practi-
tice and research methodology, and begin to develop an appreciation for the importance of communicating research results as the final stage of the scientific method.

References


Note

Send correspondence to Theresa L. White, Le Moyne College, 1419 Salt Springs Rd., Syracuse, NY 13214; e-mail: whiterl@lemoyne.edu.

A Multicomponent Approach to Teaching Sensitive Topics: Elder Abuse As an Example

Kristi S. Multhaup
Davidson College

Sensitive topics are important, if unpleasant, to discuss in psychology courses. This article describes a multicomponent approach to teaching sensitive topics like elder abuse. This approach also highlights issues of cultural influence on perceptions of (in this case) elder abuse, offers hands-on experience in interpreting data, and encourages students to identify and discuss methodological issues in research.

Even though the National Academy of Sciences (2002) reported that between 1 and 2 million Americans ages 65 or older have been mistreated by a caregiver, the teaching literature rarely addresses elder abuse. A PsychINFO search that included teach* and elder abuse yielded only three entries: One focused on nurses (Galbraith & Zdorkowski, 1984), and two focused on medical school curricula (Alpert, Tonkin, Seeherman, & Holtz, 1998; Jogerst & Hartz, 1999). By contrast, the focus of this article is using a multicomponent approach to incorporating the topic of elder abuse into an undergraduate curriculum. Although the motivation for this approach was to explore elder abuse issues, a beneficial byproduct is that the components foster discussion about how culture affects perceptions of elder abuse. Thus, this approach adds to a growing literature on teaching diversity-related issues (for a review, see Ocampo et al., 2003).

I teach about elder abuse in Psychology of Aging within the broader topic of relationships for two reasons. First, my texts (e.g., Papalia, Sterns, Feldman, & Camp, 2002; Rybash, Roodin, & Hoyer, 1995) have mentioned abuse of elderly family members. Second, I emphasize the complexity of elder adulthood by balancing the negative topic of elder abuse with more positive topics (e.g., friendships, sibling relationships). This latter point fits a course theme that elder adulthood, like any other age period, may contain a mix of positive and negative experiences and no one person experiences them all. Our week studying relationships includes two lectures on empirical data and theories related to relationships (e.g., Carstensen, Isaacowitz, & Charles, 1999) and a class period devoted to discussion of elder abuse.

In the first few weeks of class, well before our week on relationships, students complete a modified version of Moon and Williams’s (1993) survey about perceptions of elder abuse. The survey contains all 13 scenarios from Moon and Williams’s appendix and asks students whether each situation involves abuse; if so, whether the abuse is mild, moderate, or severe and whether they would ask for help if they were in each situation. The scenarios range from a spouse getting very angry and yelling over a small accident to sexual exploitation of a victim of Alzheimer’s disease.

When we study relationships, at the end of the first lecture students view two cases from the documentary A House Divided: Elder Abuse (Wright, 1989; the Elder Abuse section of American Psychological Association’s Adult Development and Aging Division Web pages describes this and other videos, http://apadiv20.pbhp.ufl.edu/vidlist.htm). The first case portrays an older Canadian woman whose son physically abused her. The second depicts an older Chinese man in San Francisco whose wife and son locked him in the basement and took his Social Security checks. The video explores community agencies designed to help older adults.

The week’s reading includes the text chapter about relationships, the Moon and Williams (1993) article, and a data summary of student responses from each time I have taught Psychology of Aging (e.g., see Table 1). The writing assignment is for students to (a) compare my Davidson student data with Moon and Williams’s data and comment on what they see as the implications of their comparison and (b) submit a comment or question for class discussion. The instructions indicate that the Affirmative Response and Yes, Seek Help columns in the table correspond to the first data columns of Tables 2 and 4, respectively, in the Moon and Williams article and that a .5 in the mild, moderate, and severe columns indicates a response was between two choices (e.g., I coded a response between moderate and severe as .5 in the moderate column and .5 in the severe column). Students submit written responses at the beginning of the class discussion.

Although discussions vary from semester to semester, several points inevitably arise. First, students discuss the issue of who decides what abuse is, particularly in the case where generational and cultural perspectives may differ—should it be well-meaning social servants or the older adult involved? Second, students note that for some of the scenarios they sympathize with the victim’s child because they know what it
Evaluation of a Brief Homework Assignment Designed to Reduce Citation Problems
Pamela Schuetze
Teaching of Psychology 2004 31: 257
DOI: 10.1207/s15328023top3104_6

The online version of this article can be found at:
http://top.sagepub.com/content/31/4/257

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Society for the Teaching of Psychology

Additional services and information for Teaching of Psychology can be found at:

Email Alerts: http://top.sagepub.com/cgi/alerts
Subscriptions: http://top.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Oct 1, 2004
What is This?
Evaluation of a Brief Homework Assignment Designed to Reduce Citation Problems

Pamela Schuetze
State University of New York College at Buffalo State College and Research Institute on Addictions
State University of New York at Buffalo

I evaluated a brief homework assignment designed to reduce citation problems in research-based term papers. Students in 2 developmental psychology classes received a brief presentation and handout defining plagiarism with tips on how to cite sources to avoid plagiarizing. In addition, students in 1 class completed 2 brief homework assignments in which they had to identify information in 1 page of text that required a citation. Results found that students who completed the homework assignments had fewer problems with citations, believed that they had a better understanding of situations that comprised plagiarism, and had more confidence in their ability to avoid plagiarism.

As many as 81% of students engage in some form of cheating at the undergraduate level (McCabe & Trevino, 1993), including cheating on tests and assignments, receiving improper assistance on assignments, and helping others cheat. One form of academic dishonesty that has received considerable attention in recent years is plagiarism (e.g., Groark, Oblinger, & Choa, 2001; Love & Simmons, 1998; Roig, 1997). Although plagiarism is not a new form of academic dishonesty, institutions throughout the United States have recently recorded significant increases in rates of plagiarism (see Groark et al., 2001). Much of the recent interest in plagiarism has stemmed from concern about the increasing prevalence of and access to term paper mills on the Internet (Groark et al., 2001). The use of term paper mills typically involves intentional plagiarism on the part of the student. Less information is available, however, on other forms of plagiarism that result from failure to cite properly paraphrased material. In many cases, failure to cite paraphrased material may be a form of inadvertent plagiarism that results from an improper or unsophisticated understanding of how to avoid plagiarism.

Although some studies suggest that the majority of students understand the distinction between plagiarism and paraphrasing (Hale, 1987; Julliard, 1994), students may not fully understand how to appropriately cite paraphrased material despite the fact that most writing manuals, including the American Psychological Association (APA) Publication Manual, discuss and exemplify proper paraphrasing and citation (e.g., Aaron, 1998; APA, 2001; Howard, 1999; Troyka, 1999). In fact, in one study, 48% of students were unaware of how to properly cite sources within the body of a research paper (Froese, Boswell, Garcia, Koehn, & Nelson, 1995). Thus, developing effective methods of educating students about how to identify material that requires a citation may reduce the rates of inadvertent plagiarism. In fact, a recent study found that students who received feedback about their performance on a paraphrasing task, examples of plagiarism, or a combination of feedback and examples were better at detecting plagiarism than students who did not get this information (Landau, Druen, & Arcuri, 2002). These results suggest that minimal intervention may increase students’ knowledge about plagiarism. However, without specific practice in detecting paraphrased information, this knowledge may not impact the ability of students to avoid inadvertent plagiarism on future writing assignments. In addition, providing feedback on students’ attempts to properly paraphrase paragraphs is a time-intensive process for the instructor and, therefore, may not be feasible in large classes or in situations where an instructor has a heavy teaching load. Thus, the purpose of this study was to evaluate the effectiveness of a brief homework assignment for psychology students to reduce the rate of plagiarism due to vague and missing citations in research-based papers.

Method

Participants

Participants included students in two life span developmental psychology classes (59 female students, 11 male students) ranging in age from 18 to 42 (M = 22.01, SD = 3.66). Both classes had the same instructor and were structured in the same way (e.g., same textbook, lecture topics, grading format). The major writing assignment for the course was to write an 8- to 10-page paper on a developmental psychology topic chosen by the student. The only sources permitted were journal articles. The instructor approved an outline of the paper, as well as a preliminary list of the students’ sources, approximately 6 weeks prior to the due date for the term paper. One class (n = 40) received a 30-min presentation and handout on APA style in-text citations and referencing, which included definitions of various forms of plagiarism, including direct plagiarism (copying information verbatim without citing or including quotation marks), inappropriately using
work from other students without giving credit to the original student, vague or incorrect citations (not clearly indicating where material from another source begins and ends), and mosaic plagiarism (superficial changes to wording or sentence structure) as well as tips on how to avoid plagiarizing. I randomly selected the other class (n = 36) to receive the same presentation and handout. In addition, they received a brief homework assignment designed to clarify the proper usage of APA style in-text citations. Both classes were unaware of their group assignment.

Homework Assignment

The homework assignments consisted of one page of text from an introduction section in a published manuscript by the author, with all citations removed. The paraphrased material could be easily identified in the chosen section (e.g., included statements such as “previous research has indicated” or “numerous researchers have”). Other paraphrased sections were somewhat harder to identify but provided good examples for class discussion. Students indicated each sentence they believed should have a citation using consecutive numbers placed at the end of the appropriate sentence. The assignment and grading criteria were explained during approximately 5 min of class. Students completed two such homework assignments over the course of the semester, each of which was worth 25 points (approximately 5% of the overall course grade).

After the first assignment, students received feedback for missing and unnecessary citations. The most mistakes were a failure to recognize that material within a paragraph could be obtained from multiple sources (44%) and inappropriately citing general statements intended to introduce a new topical section or summarize material presented in a preceding section (26%). Students then brainstormed about strategies to use to avoid repeating similar mistakes during a 5-min class discussion. They applied this knowledge while completing the second homework assignment. As expected, scores for the second homework assignment were significantly higher (M = 21.83, SD = 6.20) than scores on the first assignment (M = 18.40, SD = 7.62), paired samples t(32) = −1.96, p < .05.

To empirically evaluate that this assignment would require a minimal time commitment, students indicated how long it took to complete the assignment. Ninety-two percent of students indicated that each assignment required 15 min or less to complete; 6% indicated that it required 15 to 20 min to complete. The remaining 2% did not provide any information about the time it took to complete the assignment. Each student’s homework assignment required less than 2 min for grading and written feedback.

Results

Comparison of Course Sections

The two sections of developmental psychology did not differ in average exam grades, two-tailed t(82) = 1.66, p = .10; paper grades with grades adjusted to eliminate any penalty for citation problems, two-tailed t(82) = 1.16, p = .17; or final grades, two-tailed t(82) = 0.05, p = .96.

Assessment of Citation Problems

I used a 5-point rating scale to assess the frequency of citation problems for each term paper after grading the term paper ranging from 1 (consistent failure to cite properly throughout the paper; e.g., numerous citation failures per page) to 5 (no apparent failures to cite). A second psychology professor, who was blind to the hypothesis and group status of students, used the same scale to assess each term paper. Interrater reliability was high, r = .92. Students who completed the homework assignment had significantly fewer citation problems than students who did not complete the homework assignment, two-tailed t(73) = −4.57, p < .001. See Table 1 for means and standard deviations of frequency of citation problems for each group.

Measures of Plagiarism Knowledge

After students received the handout on types of plagiarism and heard the associated class lecture, I tested knowledge of this information by including five multiple-choice questions on a regularly scheduled exam. These questions included two factual questions to test student knowledge of the composition of types of plagiarism discussed in class, one question that required students to evaluate whether a specific piece of information should be cited, and two ques-

<table>
<thead>
<tr>
<th>Outcome Variables</th>
<th>No Homework Group</th>
<th>Homework Group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Exam average (%)</td>
<td>76.02</td>
<td>13.31</td>
</tr>
<tr>
<td>Term paper grade (%)</td>
<td>77.73</td>
<td>8.40</td>
</tr>
<tr>
<td>Final grade (out of 580 possible points)</td>
<td>461.81</td>
<td>57.13</td>
</tr>
<tr>
<td>Citation problems*</td>
<td>3.87</td>
<td>1.11</td>
</tr>
<tr>
<td>Understanding of plagiarism*</td>
<td>3.12</td>
<td>0.97</td>
</tr>
<tr>
<td>Confidence in ability to avoid plagiarism*</td>
<td>2.75</td>
<td>0.84</td>
</tr>
<tr>
<td>Understanding of importance of proper citations</td>
<td>3.90</td>
<td>0.90</td>
</tr>
<tr>
<td>Exam questions on plagiarism (maximum of 5 correct)</td>
<td>3.85</td>
<td>1.01</td>
</tr>
</tbody>
</table>

*p < .01.
ations that required students to select the appropriate paraphrased paragraph from several versions of an original paragraph. There were no differences between classes on the number of correct responses, two-tailed $t(74) = 1.20, p = .24$.

At the end of the semester, students completed a questionnaire consisting of three items designed to assess their understanding of plagiarism and their perceived ability to avoid plagiarism. They compared their perceptions of their current skill level on each item to their skill level at the beginning of the semester. They used a 5-point rating scale ranging from 1 (considerably worse) to 5 (considerably improved) to rate each item. Overall, students who completed the homework assignment indicated that they had a better understanding of what constitutes plagiarism, two-tailed $t(73) = -4.85, p < .001$, and had more confidence in their ability to avoid plagiarism, two-tailed $t(73) = -7.50, p < .001$, than the control group. There was no difference between groups in their perceived understanding of the importance of using proper citations, two-tailed $t(73) = -.96, p = .12$. See Table 1 for means and standard deviations on these items for each group.

Discussion

These results indicated that students who completed the citation homework assignments believed that they had a better understanding of situations that comprised plagiarism as well as having more confidence in their ability to avoid plagiarism. Increased student confidence in their ability to avoid plagiarism could hypothetically result in an inaccurate perception that they are fully knowledgeable about the complexities involved in proper citations in scientific papers. However, the positive nature of these perceptions was validated by less frequent citation problems observed on a research-based term paper in the group of students who completed the citation homework assignment. Thus, even a low-intensity intervention can be effective in conveying an understanding of plagiarism and reducing the frequency of citation problems. Interestingly, there was no difference between classes on exam questions designed to assess student knowledge on the definition or recognition of plagiarism. Therefore, giving students experience in identifying specific situations that require citations and providing the opportunity to discuss in class how to recognize these situations is necessary for students to apply knowledge about correctly using citations in their writing.

Although all students who participated in this study indicated that they experienced equal gains over the course of a semester in their understanding of the importance of proper citation, students who did not have an opportunity to practice and apply this knowledge showed less of an increase in their confidence and knowledge about how to avoid committing plagiarism. Thus, along with the lack of practice from doing the homework assignment, this relative uncertainty may explain the significantly higher number of citation problems in their term papers.

These results suggest two important implications. First, simply telling students what plagiarism is and how to avoid it does not provide them with the necessary skills to translate this knowledge into practice. Second, interventions that require a minimal time commitment for both students and instructors may be effective for reducing plagiarism and increasing student knowledge about and strategies to avoid plagiarism.

Although these results are promising, this study had limitations. Students were not assigned to groups randomly. Although there is no reason to believe that the two classes were different from each other in any substantial way that might impact these findings, this possibility cannot be ruled out. In addition, students who completed the homework assignment spent more time on class work (by completing the homework assignments), which may have accounted for their gains. Future studies should explore whether these gains would persist after controlling for this issue by adding lecture and demonstration in class to the control group. Despite these limitations, the results suggest that instructors can decrease levels of plagiarism by instituting fairly simple assignments and related class discussions.

References


Note

Send correspondence to Pamela Schuetze, Department of Psychology, State University of New York College at Buffalo, 1300 Elmwood Avenue, Buffalo, NY 14222–1095; e-mail: schuetp@buffalostate.edu.
An Assessment of Reliability and Validity of a Rubric for Grading APA-Style Introductions
Mark A. Stellmack, Yasmine L. Konheim-Kalkstein, Julia E. Manor, Abigail R. Massey and Julie Ann P. Schmitz
Teaching of Psychology 2009 36: 102
DOI: 10.1080/00986280902739776

The online version of this article can be found at:
http://top.sagepub.com/content/36/2/102

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Society for the Teaching of Psychology

Additional services and information for Teaching of Psychology can be found at:

Email Alerts: http://top.sagepub.com/cgi/alerts
Subscriptions: http://top.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Apr 1, 2009

What is This?
An Assessment of Reliability and Validity of a Rubric for Grading APA-Style Introductions

Mark A. Stellmack
University of Minnesota

Yasmine L. Konheim-Kalkstein
Department of Educational Psychology, University of Minnesota

Julia E. Manor, Abigail R. Massey, and Julie Ann P. Schmitz
University of Minnesota

This article describes the empirical evaluation of the reliability and validity of a grading rubric for grading APA-style introductions of undergraduate students. Levels of interrater agreement and intrarater agreement were not extremely high but were similar to values reported in the literature for comparably structured rubrics. Rank-order correlations between graders who used the rubric and an experienced instructor who ranked the papers separately and holistically provided evidence for the rubric’s validity. Although this rubric has utility as an instructional tool, the data underscore the seemingly unavoidable subjectivity inherent in grading student writing. Instructors are cautioned that merely using an explicit, carefully developed rubric does not guarantee high reliability.

Rubrics are tools for evaluating and providing guidance for students’ writing. Andrade (2005) claimed that rubrics significantly enhance the learning process by providing both students and instructors with a clear understanding of the goals of the writing assignment and the scoring criteria. Stevens and Levi (2004) noted that rubrics facilitate timely and meaningful feedback to students. Peat (2006) suggested that, because of their explicitly defined criteria, rubrics lead to increased objectivity in the assessment of writing. Thus, different instructors might use a common rubric across courses and course sections to ensure consistent measurement of students’ performance. To this end, this article describes the assessment of the reliability and validity of a rubric that was designed for use in grading APA-style introductions in multiple sections of an introductory research methods course.

Rubrics used in many subject areas in higher education generally include two elements: (a) a statement of criteria to be evaluated, and (b) an appropriate and relevant scoring system (Peat, 2006). Rubrics can be classified as either holistic or analytic (Moskal, 2000). Holistic rubrics award a single score based on the student’s overall performance, whereas analytic rubrics give multiple scores along several dimensions. In analytic rubrics, the scores for each dimension can be summed for the final grade. Although an advantage of the holistic rubric is that papers can be scored quickly, the analytic rubric provides more detailed feedback for the student and increases consistency between graders (Zimmero, 2004).

Regardless of its format, when used as the basis of evaluating student performance, a rubric is a type of measurement instrument and, as such, it is important that the rubric exhibits reliability (i.e., consistency of scores across repeated measurements) and validity (i.e., the extent to which scores truly reflect the underlying variable of interest). Although reliability and validity have been noted as issues of concern in rubric development (Moskal & Leydens, 2000; Thaler, Kazemi, & Huscher, 2009), the reliability and validity of grading rubrics seldom have been assessed, most likely due to the effort and time commitment that is required to do so.

Two types of reliability that can be evaluated for a grading rubric are intrarater agreement, the extent to
which scores resulting from use of the rubric are consistent across graders, and interrater reliability, the correlation between the scores of different graders (Tinsley & Weiss, 2000). When the task of grading students’ papers is divided among multiple graders, as is the case in our course and in many undergraduate courses at large universities, high interrater agreement is particularly important to achieve uniform grading across sections of the course. (Thaler et al., 2009, in contrast to this study, evaluated their rubric in terms of interrater reliability.) For the remainder of this article, when we refer to the general concept of reliability, we do so with respect to the computation of interrater agreement.

One way in which the interrater agreement of a rubric can be enhanced is through the careful description of criteria on which grades will be based. Without specific grading instructions, one grader might emphasize grammar, for example, whereas another might emphasize content. A well-developed rubric guides graders in placing the desired emphasis on specific, uniform criteria, so that the role of subjective opinions is minimized (Newell, Dahn, & Newell, 2002). Training of graders usually is necessary to maximize interrater agreement (Zimmaro, 2004). A reliable measuring instrument also is one that yields the same score in repeated measurements. With respect to a grading rubric, this task amounts to showing that the same grader would assign the same grade to a given paper if the grader were to grade the paper again. The repeatability of grades within graders is called intrarater agreement (Moskal & Leydens, 2000).

To say that a rubric exhibits validity means that it measures the underlying variable of interest. In this case, the variable of interest is “quality of psychological writing.” One way to demonstrate validity is to provide evidence that different measures of the same variable are correlated with one another (i.e., convergent validity; Crocker & Algina, 1986). In this case, we assessed validity by comparing measures of writing quality obtained with the rubric to judgments of writing quality made by an independent evaluator who has experience in teaching and grading student writing but who did not use the rubric.

The purpose of this study was to develop a rubric for grading student writing of the introduction section of an APA-style manuscript and to evaluate the rubric’s reliability and validity. We chose to focus on the introduction section because those involved in teaching our research methods course frequently identify it as the most difficult assignment in the course for instructors to grade and for students to write. In addition, we chose to analyze a rubric for only the introduction rather than for an entire manuscript in an attempt to simplify our task of maximizing interrater and intrarater agreement. In this article, we briefly describe the development of an analytic grading rubric, present data regarding interrater agreement and intrarater agreement in the use of the rubric, and assess the rubric’s validity.

Method

Writing Assignment and Graders

In a section of the research methods course at the University of Minnesota, students designed and conducted a research project of their own choosing in groups of approximately 4 students. The students were given a writing assignment in which they were to compose a well-written APA-style introduction, including a literature review of at least five peer-reviewed sources, in order to build an organized argument leading to a statement of the hypothesis of the students’ research project. Each student wrote his or her own paper (rather than a common group paper).

The researchers (the five authors) who developed and evaluated the rubric in this study included two instructors who had taught the research methods course during four previous semesters, a graduate student who had acted as a teaching assistant in the research methods course during one semester, and two advanced undergraduates who had taken the research methods course and who had been identified by their instructors as exceptional students in the course. In the research methods course at the University of Minnesota, student papers are usually graded by first- or second-year graduate student teaching assistants. As such, we felt that it was appropriate to include instructors as well as advanced undergraduate students who had demonstrated high levels of writing ability.

Rubric Development

In the early stages of developing a rubric, we began with an extremely detailed rubric that had been used in the past and that contained more than 30 specific statements that could be evaluated with respect to the student’s paper (e.g., “You did not indicate how your sources relate to your study” and “You did not integrate your sources into a coherent whole argument”). The specificity of the statements coupled with their inherent subjectivity led to frequent disagreements among graders as to whether a particular criterion was satisfied.
Table 1. Brief Descriptions of the Eight Dimensions in the Rubric

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Quality Assessed</th>
</tr>
</thead>
<tbody>
<tr>
<td>APA formatting</td>
<td>How well the paper follows the rules of APA formatting</td>
</tr>
<tr>
<td>Literature review and argument support</td>
<td>Description of previous literature and its application to the current study</td>
</tr>
<tr>
<td>Purpose of study</td>
<td>How well the student clarified why the topic should be researched, as well as how student's study follows from previous research</td>
</tr>
<tr>
<td>Study description and hypothesis</td>
<td>Student's brief description of what would be done in his or her study and the hypothesis</td>
</tr>
<tr>
<td>Overall organization and logical flow</td>
<td>How well the paper builds a coherent, smoothly flowing argument that culminates in the hypothesis</td>
</tr>
<tr>
<td>Sources</td>
<td>The quality of sources used, including whether the student used at least five peer-reviewed sources that are relevant to the topic of the study</td>
</tr>
<tr>
<td>Scientific writing style</td>
<td>Whether the paper uses appropriate scientific language and style</td>
</tr>
<tr>
<td>Composition/grammar/word choice</td>
<td>How well the paper is written in terms of grammar, sentence structure, and phrasing</td>
</tr>
</tbody>
</table>

As a result, we discovered, as have others (Peat, 2006; Stevens & Levi, 2004; Thaler et al., 2009), that a smaller number of criteria was more practical.

We reduced the grading criteria to eight broad dimensions of content that were of particular importance for this type of writing assignment (see Table 1), each containing four possible scoring levels (0–3 points). It is generally considered to be advantageous to have fewer scoring levels with meaningful distinctions than to have more scoring levels where it might be difficult to distinguish between categories (Moskal, 2000; Stevens & Levi, 2004). We decided that four scoring levels for each dimension gave the desired resolution to distinguish between levels of achievement. (See Table 2 for an example of the scoring levels for one dimension.)

Each week, during the course of about 10 weeks, we independently graded several introductions written by research methods students and convened to examine our scores. We discussed disagreements and made adjustments to the rubric in a way that we felt would prevent those disagreements in the future. Adjustments involved altering the wording that defined the dimensions and that distinguished between scoring levels within a dimension. We repeatedly refined the rubric in this way until we believed that we achieved maximum interrater agreement. (The complete, final version of the rubric that was evaluated in this article is available online at http://www.psych.umn.edu/psylabs/acoustic/rubrics.htm or via e-mail from the first author.)

Evaluating Interrater Agreement

Forty papers were selected randomly from a pool of papers written by students in the research methods class. All identifying information was removed, and none of the researchers knew the final scores that were assigned to the papers in the course. Each researcher graded 24 of the 40 papers, such that three researchers graded each paper. Scores were then compared and interrater agreement was calculated based on the number of agreements for each dimension for each paper. Thus, there were 320 potential opportunities for agreement across all papers (40 papers × 8 dimensions).

Two estimates of agreement, conservative and liberal, were calculated. Agreement was defined conservatively as all scores assigned for a dimension being equal across the three graders of that paper. Agreement was defined liberally as all scores assigned for a dimension by the three graders being within 1 point of one another. These criteria are accepted in the measurement literature (Tinsley & Weiss, 2000) and have been applied in past studies of interrater agreement for grading rubrics (Newell et al., 2002). Agreement on total overall score out of 24 possible points (8 dimensions × 3 points maximum for each) for the 40 papers was also calculated and is described in the results.

Evaluating Intrarater Agreement

Approximately 2 weeks after grading the papers, each grader regraded five of the papers that he or she originally graded. All five graders regraded different papers from one another, such that 25 different papers were regraded. To compare the first grading of each paper to the second grading, intrarater agreement was evaluated quantitatively in the same way as interrater agreement described earlier. A total of 25 papers were thus examined to assess intrarater agreement.
Table 2. An Example of the Grade Levels Within a Dimension of the Rubric

<table>
<thead>
<tr>
<th></th>
<th>3 Points</th>
<th>2 Points</th>
<th>1 Point</th>
<th>0 Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study description and hypothesis</td>
<td>• The paper gives a general description of what the research will entail (what will be done) without exhaustive methodological details.</td>
<td>• The study description and hypothesis are present but one or both are unclear.</td>
<td>Either the study description or hypothesis is missing.</td>
<td>The study description and hypothesis are missing.</td>
</tr>
</tbody>
</table>

Evaluating Validity

To determine whether the rubric truly measured the intended variable “quality of psychological writing,” we examined whether scores assigned using the rubric aligned with an experienced psychology instructor’s ranking of those same papers. The average score (out of 24 possible points) was computed across the three graders for each of the 40 papers. Ten papers were chosen from approximately equally spaced intervals across the range of scores, based on the average score. An instructor of another methods-related psychology course at the University of Minnesota who requires APA-style writing was asked to rank order the 10 papers based on a holistic judgment of their content and quality with respect to what she would consider well-written APA-style introductions. The instructor was given only the description of the writing assignment that was given to the students, and the instructor was asked to apply whatever criteria she felt was appropriate in rank ordering the papers. The Spearman rank-order correlation coefficient between the independent judge’s rankings and the rankings based on the average scores of the papers using the rubric was computed.

Results

Interrater agreement by the liberal definition (i.e., with all graders scoring within 1 point for a given dimension) was 287/320 = .90 (Cohen’s kappa = .33). Chance agreement given four scoring levels and three graders would be .34 and .06 by the liberal and conservative definitions, respectively. Newell et al. (2002) found comparable levels of agreement for three graders using a rubric for grading students’ solutions of chemical engineering problems, a task that was not writing based. The rubric developed by Newell et al. also had four scoring levels within each dimension. Newell et al. reported interrater agreement liberally defined of .93 and conservatively defined of .47.

Analysis of the total score assigned to each paper out of 24 points showed that all graders arrived at the same total score for 5% of the papers. The total scores on 75% of the papers were within 4 points of each other. The maximum range of total scores for a single paper was 8, observed for only one paper.

Intrarater agreement (i.e., when the same papers were graded a second time by the same graders) was 196/200 = .98 by the liberal definition and 156/200 = .78 by the conservative definition. Given that each paper was graded twice, chance agreement would be .62 and .25 by the liberal and conservative definitions, respectively.

In evaluating the validity of the rubric, the Spearman rank-order correlation coefficient between the independent judge’s rankings and those obtained with the rubric was .49. For comparison, the mean Spearman rank-order correlation coefficient between rankings of the five graders who scored papers with the rubric was .54. Thus, the correlation between rankings of those who used the rubric and the judge who applied her own independent criteria is comparable to the correlation between graders who used the rubric and the same explicitly stated criteria. This result provides support for the validity of the rubric.
Discussion

Although rubrics are used frequently in evaluating student writing, little research has focused on assessing the quality of rubrics as measurement instruments (i.e., their reliability and validity). Clearly, it is desirable to establish that a rubric is a valid measure of the variable that one is attempting to evaluate. The rubric’s reliability is of particular concern to instructors of large psychology classes that use multiple graders.

We undertook development of the present rubric with the optimistic view that near-perfect interrater agreement could be obtained through careful and diligent refinement of the rubric. That turned out not to be the case, with graders reaching perfect agreement only 37% of the time, and graders agreeing perfectly with themselves (intrarater agreement) only 78% of the time. Note that development and evaluation of the rubric spanned much of a semester, which probably represents a greater degree of effort than most instructors typically devote to rubric development. Although interrater agreement and intrarater agreement were not as high as we might have hoped, the rubric exhibited a reasonable degree of reliability in that three graders agreed with each other within 1 point 90% of the time. In addition, as indicated earlier, the introduction is frequently identified as the most problematic writing assignment for both writers and graders in the research methods course. Therefore, we might have expected it to be particularly difficult to attain high reliability for a rubric designed for use in grading introductions. As a whole, these data establish important benchmarks in the evaluation of grading rubrics, primarily because of the lack of such data.

Interpreting reliability and validity coefficients as “high” or “low” is a subjective proposition that is dependent on the way in which the information provided by the coefficients will be used and, as such, is of limited practical use. As long as the coefficients exceed chance, as in this study, reliability and validity exist to some degree. For our purposes, the results of this study suggest ways in which the numerical scores obtained using the rubric can be converted to letter grades. For example, given that numerical scores spanned the entire range of possible scores from 0 to 24, it is reasonable for the letter grade categories to be well distributed across the range of possible scores. Furthermore, given that the rubric used here resulted in graders assigning point totals within 4 points of one another on 75% of the papers, one might consider making the range of each letter grade for this writing assignment at least 4 points. For example, in our application of the grading rubric, after considering the overall quality of writing that existed in different ranges of scores, we intend to assign a grade of A for scores from 21 to 24 and a grade of F for scores from 0 to 3 (out of 24).

The fact that the graded introductions were based on student-designed studies might have influenced the reliability of the grading. In many research methods courses, students write their first APA-style manuscripts describing a research project that the instructor provides. It might be expected that an introduction describing a canned experiment would produce greater reliability because the ideal content and structure of the introduction could be more specifically defined for the graders.

The rubric used here was developed over the course of approximately 10 weeks during which the graders became increasingly familiar with its use and with each other’s grading tendencies. In effect, the development period also served as a training period for the graders. Even after such an extensive period of training, the interrater agreement remained imperfect. The training period in most real-world situations likely would amount to one or two sessions at the beginning of a semester. As a result, one might expect interrater agreement to be lower in practice than that obtained here, particularly when there is substantial turnover among graders from semester to semester as in our research methods course.

The results of this study underscore the inherent subjectivity of evaluating student writing. This subjectivity is problematic if one desires a grading rubric that can produce objective assessments across graders and course sections. An understanding of the reliability of a rubric can aid an instructor in converting scores obtained with the rubric into letter grades by revealing the potential variability associated with assigning a score to any particular paper. The fact that our rubric displayed unexpectedly low interrater and intrarater agreement will lead us to reconsider the way in which student writing is scored in our research methods course and the way in which those scores are incorporated into final grades. At the very least, we hope that this study will lead others to more rigorously assess grading rubrics as measurement instruments.

References

Notes

1. We thank Dr. Gail Peterson and Jamie Peterson for their assistance with this project. We also thank Dr. Randolph Smith and three anonymous reviewers for providing valuable feedback on an earlier version of this manuscript.
2. Yasmine L. Konheim-Kalkstein is now a Psychology faculty member at North Hennepin Community College.
3. The rubric that we evaluated in this article is available online at http://www.psych.umn.edu/psylabs/ acoustic/rubrics.htm or via e-mail from the first author.
4. Send correspondence to Mark A. Stellmack, Department of Psychology, University of Minnesota, 75 East River Pkwy., Minneapolis, MN 55455; e-mail: stell006@umn.edu.
Teaching Scientific Writing: Measuring Student Learning in an Intensive APA Skills Course
Vickie R. Luttrell, Jana L. Bufkin, Valerie J. Eastman and Robin Miller

Teaching of Psychology 2010 37: 193
DOI: 10.1080/00986283.2010.488531

The online version of this article can be found at:
http://top.sagepub.com/content/37/3/193

Published by:
http://www.sagepublications.com

On behalf of:
Society for the Teaching of Psychology

Additional services and information for Teaching of Psychology can be found at:

Email Alerts: http://top.sagepub.com/cgi/alerts
Subscriptions: http://top.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Jul 1, 2010
What is This?
Teaching Scientific Writing: Measuring Student Learning in an Intensive APA Skills Course

Vickie R. Luttrell, Jana L. Bufkin, Valerie J. Eastman, and Robin Miller
Drury University

The Publication Manual of the American Psychological Association (American Psychological Association [APA], 2010) provides a widely utilized template for preparing research reports in the behavioral sciences. Because learning APA style is integral to disciplinary socialization, our department recently implemented a 1-hr course in scientific writing. Using a pre- and posttest factorial design, we compare changes in APA knowledge over an academic semester and quality of final papers in 2 groups of students: those enrolled in scientific writing (n = 20) and others who were expected to master APA basics on their own (n = 18). Results suggest that intensive instruction in APA style is beneficial for student learning outcomes.

Discipline-specific writing classes vary in design, from the introduction of writing exercises within existing classes (Fallahi, Wood, Austad, & Fallahi, 2006; Nadelman, 1990; Poe, 1990; Ware, Badura, & Davis, 2002) to courses devoted solely to scientific writing (Goddard, 2003; Pfeifer & Ferree, 2006; Rileigh, 1998), but a common objective is to enhance student familiarity with the Publication Manual of the American Psychological Association (American Psychological Association [APA], 2010). Students are required to use the APA manual in classes because it provides the most widely used template for scientific thinking and writing in the behavioral and social sciences and is, therefore, integral to students’ professional socialization (Ballard, Klein, & Dean, 2007; Gianaros, 2006; Madigan, Johnson, & Linton, 1995).

Strategies for advancing professional socialization through scientific writing are more plentiful than evaluations of their effectiveness, but evidence suggests that students who receive instruction on APA writing requirements are better equipped to be scientific writers than other students (Fallahi et al., 2006; Goddard, 2003; Poe, 1990; Rileigh, 1998). Given the importance of scientific writing to professional socialization and the potential for improving students’ understanding of APA guidelines through intensive instruction, our department developed a class called Scientific Writing that was implemented in spring 2007. Scientific Writing is required of all criminology, psychology, and sociology majors and serves as a prerequisite to courses in research methods, statistics, and undergraduate research. Four sections of the course are offered each year with enrollments capped at 15 students per class.

Our course is akin to other classes created for the purpose of improving students’ scientific writing skills. Similar to courses presented by Pfeifer and Ferree (2006) and Rileigh (1998), for example, Scientific Writing is designed for students who have no exposure to APA guidelines. Similar to Goddard’s (2003) and Rileigh’s courses, the emphasis of the class extends beyond APA referencing. Elements emphasized include database searches, title pages and abstracts, reduction of language bias, in-text citations, numbers and abbreviations, and reference lists. Yet, Scientific Writing is different from those classes in that it is a 1-hr rather than a 3-hr course, which we view as advantageous both in terms of the time commitment involved and associated tuition costs. We evaluated Scientific Writing in spring 2008, and our design represents an improvement over other studies due to the use of a comparison group, which allows for a determination of whether intensive instruction is more effective than self-mastery of APA guidelines.
Method

Participants

Twenty students across two sections of Scientific Writing comprised the experimental group. Those participants ranged in age from 18 to 22 years (M = 19.60, SD = 1.14), attended school full time (100%), and were predominantly White (95%) and women (70%). Most were first- (45%) or second-year (35%), non-transfer (85%) students. The comparison group was comprised of students enrolled in a higher level psychology course (N = 18), who ranged in age from 18 to 22 years (M = 20.50, SD = 1.15). These students attended school full time (100%) and were also predominantly White (89%), women (78%), and nontransfer (89%) students. Sixty-seven percent were juniors or seniors. No students in the comparison group had completed Scientific Writing; however, about 28% had already taken research methods and statistics.

Measures and Procedures

Student mastery of APA guidelines was assessed through scores obtained on a 22-item multiple-choice questionnaire adapted from Levy and Tartaro (2007). After obtaining informed consent, participants in both groups were pretested on the first day of class. Posttests were administered at the end of the semester. We also assessed student performance on a required literature review that was graded on a 100-point scale and included a title page, an abstract, seven to eight content pages, and a reference list. Because most points associated with the literature reviews involved specific APA issues, there was little room for instructor subjectivity in grading. However, to help ensure consistency of grading, instructors adopted common grading rubrics across courses.

Results

We scored items on the APA competency questionnaire on a pass–fail basis. Given that test items seemed to cluster around three general themes (i.e., reference list issues [nine items], in-text citations [six items], and miscellaneous issues [seven items]), we applied the Kuder–Richardson 20 (KR20) formula to pre- and posttest scores to determine if we could combine students’ responses into three subscale scores. KR20 estimates for scores at the subscale level were generally low, ranging from .05 to .58. KR20 estimates for the total-scale scores at pre- and posttest were $\alpha = .68$ and $\alpha = .71$, respectively. Consequently, we decided to examine performance on the questionnaire as a whole.

At pretest, participants in the experimental group achieved scores ranging from 6 to 19 (M = 11.30, SD = 3.56); their posttest scores ranged from 10 to 22 (M = 18.50, SD = 2.82). Scores obtained by students in the control group ranged from 5 to 20 (M = 12.83, SD = 3.65) at pretest and from 10 to 20 (M = 16.28, SD = 2.95) at posttest. To compare knowledge gains between groups, a 2 (intensive instruction vs. self-mastery) $\times$ 2 (pretest vs. posttest) factorial analysis was performed. The analysis revealed a main effect for time, with students in both classes experiencing statistically significant gains in APA competence, $F(1, 36) = 85.56, p < .001, \eta^2_p = .70$. The interaction term was also statistically significant, with students in the course experiencing greater increases in APA competence over time than comparison students, $F(1, 36) = 10.65, p = .002, \eta^2_p = .23$. A breakdown of this interaction showed that although performance differences in APA competence were not statistically different at baseline, $t(36) = -1.31, p = .198, d = -.42$, by the end of the semester, Scientific Writing students outperformed students in the control group, $t(36) = 2.38, p = .023, d = .77$.

Students’ literature reviews ranged from 57 to 98 points (M = 84.18, SD = 9.96). An independent-samples $t$ test revealed that differences in paper writing performance between Scientific Writing students (M = 84.30, SD = 10.94) and students in the control group (M = 84.06, SD = 9.05) were not statistically significant, $t(36) = .08, p < .94, d = .02$.

Discussion

Our evaluation of Scientific Writing differed from past studies examining the effectiveness of teaching APA guidelines (Fallahi et al., 2006; Goddard, 2003; Poe, 1990; Rileigh, 1998) because we were able to utilize a quasi-experimental design with a comparison group. Although both groups experienced statistically significant gains in APA competency at posttest, students in the 1-hr APA intensive course demonstrated statistically greater advances in APA knowledge than students in the (self-mastery) control group.
Given that this knowledge did not translate into statistically significant differences in paper writing performance and, as noted, the control group improved in APA competency as well, some might argue that self-mastery is as effective as intensive instruction. However, we believe this conclusion is unlikely due to potential advantages held by the comparison group. Compared to the Scientific Writing group, students in the control group were older, on average, $t(36) = -2.42$, $p = .021$, $d = .79$, and more likely to be juniors or seniors, $\chi^2(1, N = 38) = 8.46$, $p = .004$. Consequently, it is likely that many of the control participants had already written APA-style papers or were at least somewhat familiar with APA formatting issues prior to enrollment in the comparison course. It is also possible that our failure to find a group difference lies in grading variations across courses, a possibility that was minimized, however, by the use of a common grading rubric.

Clearly, the comparison group was imperfect, and both groups were small because students in our department were transitioning between old and new curricula. Nonetheless, we are encouraged by Scientific Writing’s ability to generate an APA competency that is above and beyond students’ capacity when learning APA techniques on their own. We are also encouraged that first- and second-year students perform on par with more experienced students when compiling literature reviews. Thus, we believe the findings presented strengthen the case for inclusion of courses aimed at building APA competency in undergraduate curricula. Moreover, evidence suggests that gains in scientific writing ability can be achieved in 1-hr courses, which are less time and resource intensive than traditional 3-hr approaches.

References


Notes

1. This material is based on work supported by the National Science Foundation under Grant DUE-0536161.

2. The Scientific Writing class is part of a larger curricular infrastructure known as the Scientific Core. The Scientific Core, which is required of all behavioral science majors, includes courses in scientific writing (1 hr), research methods (3 hr), and statistics with laboratory (4 hr) as well as a two-semester, team-based research experience.

3. Send correspondence to Vickie R. Luttrell, Department of Behavioral Sciences, Drury University, Springfield, MO 65802; e-mail: vluttre@drury.edu.
Guided Experimentation with Databases Improves Argumentative Writing
Phanikiran Radhakrishnan, Dianne Lam and Elaine Kong Tamura
Teaching of Psychology 2010 37: 210
DOI: 10.1080/00986283.2010.488547
The online version of this article can be found at:
http://top.sagepub.com/content/37/3/210

Published by:
http://www.sagepublications.com

On behalf of:
Society for the Teaching of Psychology

Additional services and information for Teaching of Psychology can be found at:

Email Alerts: http://top.sagepub.com/cgi/alerts
Subscriptions: http://top.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Jul 1, 2010
What is This?
Guided Experimentation With Databases Improves Argumentative Writing

Phanikiran Radhakrishnan, Dianne Lam, and Elaine Kong Tamura
University of Toronto at Scarborough

The guided experimentation method requires that students learn how a database works by generating and testing hypotheses about it (i.e., become skilled at strategies of experimentation). These strategies are essential to learning by inquiry, which is linked to argumentation. We conducted a quasi-experimental field study to show that students who experimented with a library database in a structured manner \( (n = 60) \) wrote better argumentative papers than those who did not \( (n = 29) \). Future research should examine whether those who approach a task with the aim of learning it benefit more from opportunities to learn from experimentation with it.

Students across disciplines need to learn the fundamentals of argumentation, which involves supporting a hypothesis with evidence and using logic to explain it. Toward helping students find appropriate evidence, teachers have begun to include information literacy activities in the curriculum (Cameron & Hart, 1992; McCarthy & Pusateri, 2006; Poe, 1990). Such efforts might help students find relevant evidence (Wang, 2006), but might not improve their argumentation (cf. Berner et al., 2002) if teachers fail to focus on all aspects of it, namely, how to use logic to explain the evidence. We wondered whether teachers can improve argumentative skills with information literacy activities that focus on all aspects of argumentation. We conducted a field study to test the effectiveness of such activities.

The information literacy activity we used was the “guided exploration” method of teaching students how library databases work (Debowski, Wood, & Bandura, 2001). The approach is “guided” because the instructor directs the steps to take to learn about the relevant features of the database and sequences the order in which the steps are learned. Exploration (i.e., experimentation) is accomplished by providing time between steps for students to test hypotheses about how the database works on their own. Periods of student experimentation are interspersed with periods of instruction.

Previous research demonstrates the effectiveness of the guided exploration method on learning outcomes. For example, those who explore a computer program or a library database in a structured manner learn it better than those who explore it in an unstructured manner (Debowski et al., 2001; Van Der Linden, Sonnentag, Frese, & Van Dyck, 2001). Trudel and Payne (1995) found that structuring experimentation to one aspect of a device at a time resulted in better learning. Such manipulations enable participants to understand how something works by generating and testing their hypotheses, evaluating the results of their trials, reflecting on their actions, and changing their behavior accordingly (Van Der Linden et al., 2001).

The features of guided exploration are essential to learning by inquiry. Inquiry involves argumentation because it entails generating and assessing hypotheses via experimentation (Hillocks, 1982, 1995). Therefore, we proposed that teaching students to explore how a library database works in a guided manner should also teach them strategies of inquiry, which in turn should also improve their argumentation. This line of reasoning explains why efforts to teach students how a database works without teaching them how to experiment with it do not improve skills of inquiry (Berner et al., 2002), whereas those that incorporate experimentation do (Maor, 1991).

We conducted a quasi-experimental field study by comparing the argumentative writing of students enrolled in a course during semesters in which the information literacy activities emphasized guided exploration with those enrolled in semesters during which such activities did not emphasize exploration. Typically, students enrolled in the course were exposed...
to information literacy workshops. However, we wanted to see if emphasizing experimentation with the database in these workshops during some semesters could improve writing. We predicted that those who participated in workshops emphasizing guided exploration would write significantly better argumentative papers than those who did not.

Method

Participants

Participants were university students enrolled in an organizational behavior course over three semesters during which we collected data. Students enrolled in the latter two semesters received an information literacy workshop that emphasized guided exploration (experimental semester), whereas those in the first semester received a workshop that did not (control semester). Students took this course as part of the requirements of their major. They were all required to write an argumentative term paper as part of 20% of their course grade. We did not collect identifying information such as gender or age because asking for it was not part of the students’ regular course activities and not pertinent to our hypotheses. We asked for students’ consent to use their coursework for research. All students’ papers were photocopied before they were graded by the teaching assistant for the course and these copies were used for this study.

Based on an analysis of the power needed to detect for the effects of semester and the experimental manipulation, we selected 30 papers from each semester. We then selected an equal number of papers from each of the multiple sections taught by the same instructor within each of the three semesters. This resulted in an overall sample size of 89 papers, with 60 papers from the semesters that involved the experimental manipulation and 29 from the semesters receiving the control condition (one of the randomly selected papers from the control did not have any identifying information so we could not use it in our analyses).

Measures

We measured the efficacy of the guided method of teaching students how to explore a library database on argumentative writing. The definition of argumentation used was based on McCann (1989) and Knudson’s (1992) adaptation of Toulmin’s (1958) model. This model involves claims, data, warrants, proposition, recognition of opposition, and response to opposition. The terms used in this model were contextualized to better help students understand the nature of argumentation (cf. Gleason, 1999). Toward this end, students were given a sheet of guidelines specifying that their papers should (a) have a hypothesis with clearly defined variables, (b) describe relevant evidence supporting and countering it, (c) explain how the evidence did and did not support the evidence, and (d) explain why there was inconsistent evidence (cf. Hillocks, 1986, 1995).

Argumentative papers were scored on seven dimensions: descriptive and summarization skills, quality of resources used for definition, definitional skills, quality of resources as supporting and countering evidence for hypothesis, quality of resources used for explaining supporting and countering evidence, quality of resources to reconcile supporting and countering evidence, and argumentation skills (see Appendix for details).

The rater, who was not the teaching assistant for the course, was blind to the conditions of each paper and the papers were randomly ordered during scoring. The papers did not have any identifying information (e.g., name or semester information) other than the student number on them. The rater scored the papers only after the teaching assistant had graded all the papers from all the semesters, the paper grades were incorporated into each student’s final course grades, and these grades were submitted to the university registrar.

Before scoring the papers involved in testing the hypotheses for this study, the rater was trained by the instructor. For this, the instructor and the rater (a graduate student who is not the co-author of this paper) first discussed the meaning of the dimensions used to judge the quality of the papers. Then, they independently scored a set of papers that were not part of the set that had been selected to be part of the main analyses. Next, we computed inter-rater reliability between these scores. The rater and the instructor then discussed the cause of discrepancies in instructor–rater scoring. Subsequently, they independently scored more papers that were still not part of the study and then we computed the intrarater reliability again and there was further discussion between the rater and the instructor. When the rater and the instructor reached an acceptable level of agreement (i.e., inter-rater reliability = .70), the rater then independently scored the papers that had been randomly selected to be scored for the main analyses of the study (as was described in the “Participants” section).
Procedure

A week after students received written guidelines on the requirements of their term papers, they attended the information literacy workshop. They were first informed of the relevance of the information literacy activity for their term paper completion and that the workshop was geared toward teaching them how to search an electronic database that contained relevant articles they could use to write their papers. The workshop was conducted at the computer facilities on the university campus. Each student was seated at a separate computer that was connected via the Internet to the library database.

The instructor explained that once a research question was identified, the first step was to break down the question into key concepts and identify synonyms of these concepts. Then the instructor demonstrated how he or she would do this first step using the question, “Does how happy you are with your job affect how devoted you are to your company?” identifying the concepts of “job satisfaction” and “organizational commitment.” We then gave students an opportunity to do this first step for their preassigned research question (e.g., “Does the amount of pressure you feel at work affect how long you will be at the company?”). The instructor next demonstrated the second step by showing students (with the instructor’s research question) how to identify technical terms that are used by the database to describe each of the concepts in the research question (e.g., for job satisfaction the terms would be “attitudes of employees” and “employee happiness”). Again, the students were given time to do the second step for one of the concepts they generated from breaking down their research question. Then the instructor demonstrated how to link all the technical terms used to describe one of the concepts with the “or” Boolean operator to identify all articles using that set of technical terms. The fourth step was to keep repeating the first three steps until all technical terms and relevant articles were found for each concept in the research question. For example, in the instructor’s research question, the task was to repeat the steps for “organizational commitment.” The next step was to link technical terms used for each concept with the “and” Boolean operator to find common articles using all the concepts in the research question. The final step was to evaluate the type of articles found and to assess whether different concepts were needed to refine the type of articles found.

Students in the control group were also first informed of the relevance of the information literacy workshop. However, a key difference between the workshop delivered to the experimental group and that delivered to the control group was that the latter workshop did not allow for experimentation. In the control group’s workshop, the instructor demonstrated to the students how he or she would go about searching for articles on the instructor’s research question. However, unlike the experimental group, control group students were not given an opportunity to experiment in a guided manner with a different preassigned research question. There was no systematic breakdown of the instructional components into manageable steps and no systematic pauses between the steps to allow for students to interact or experiment with the database. The students could choose if they wished to imitate the instructor during the demonstration period or conduct their own search using their own research question (which might have varied in specificity). To summarize, all students in the study received an information literacy workshop of the same duration; however, the nature of the instruction during the workshop was different for those in the experimental versus the control group: One was encouraged to experiment in a guided manner, whereas the other was free to do as it wished.

Results

The rater scored each paper on seven dimensions and could assign a score from 1 to 10 for each dimension. The actual minimum score was .29 and the maximum was 8.5. Because the inter item reliability across the seven dimensions was very high (.94), we collapsed the ratings for each item and report the overall means for each paper.

As predicted, students who participated in information literacy activities emphasizing guided exploration wrote significantly better argumentative papers (M = 5.4, SD = 1.9) than those who received information literacy activities that did not (M = 4.4, SD = 2.0), t(87) = −2.35, p < .05, d = .51. There were no differences between the sections within the experimental and control groups.

Discussion

Students improved their argumentative writing because we provided them with information literacy activities that allowed for guided exploration. The
activities emphasized how to generate and test hypotheses and to make logical inferences by observing the results. Specifically, students in our study learned how a library database worked by experimenting with it in a structured manner. The information literacy activities taught students how to use the strategies of inquiry (Maor, 1991), which involves argumentation (Hillocks, 1995). Thus, students who engaged in guided exploration wrote better argumentative papers than those who did not.

Much of the research on the effects of information literacy activities examine if students can find evidence relevant to their hypotheses (e.g., Wang, 2006) and not if they can use logic to explain such evidence (see Scharf, Elliot, Huey, Briller, & Joshi, 2007, for an exception). Our study addresses this gap by showing that information literacy activities focusing on experimentation can predict all aspects of argumentative writing.

Additionally, although Scharf et al.'s (2007) research examined the relation between information literacy and writing, it did not manipulate the type of information literacy activities that students engaged in, nor did it assess the nature of such activities. It also does not provide complete information on the nature of writing measured. This study seemed to assess the quality of written portfolios using subjective judgments of argumentation. Our study addresses these shortcomings. We measured the effectiveness of our information literacy activities (i.e., the guided exploration method) by operationalizing argumentation skills in terms of how to make claims and support them with data.

The guided exploration method that we used to teach information literacy also enables us to reach the standards of information literacy of the Association for College & Research Libraries (2000). These standards include determining the extent of information needed, evaluating information and its sources critically, incorporating selected information into one’s knowledge base, and using information effectively to accomplish a specific purpose. Such standards can only be achieved by students who are well versed in argumentation. This assumption is supported in part by Scharf et al. (2007), who found that students who used relevant, good quality, accurate, and authoritative sources to build a foundation, compare, contrast, and refute arguments, and to synthesize ideas to construct new concepts were also the ones who wrote better papers.

Our study has several strengths. It is one of the first to show the positive effects of the guided exploration method in a naturalized setting, further attesting to the power of this pedagogical technique. Previous studies using this method (Debowski et al., 2001; Trudel & Payne, 1995; Van Der Linden et al., 2001) were conducted in laboratory settings. We illustrate the generalizability of the effect by demonstrating it in a classroom with students who were all highly engaged with the learning task so that they could do well on their term papers, which were worth 20% of their course grades.

Van Der Linden et al. (2001) speculated that structured experimentation enables participants to understand how something works by generating and testing their hypotheses and by evaluating the results of their trials. We further extend this research by showing that engaging students in structured experimentation activities goes beyond teaching them how something works. It can improve their understanding of experimentation and consequently, argumentation.

Our participants were students who had to take this course as part of their major. This allowed us to circumvent the self-selection bias and examine the effects of the method on a broad spectrum of ability and skill level. We also controlled for instructor effects by having the same instructor teach all sections of the experimental and control semesters. However, the instructor was not blind to the conditions. We attempted to control for expectancy effects by having a rater who was blind to the conditions rate the papers in a random order. Finally, because the study was part of an effort to improve information literacy pedagogy, we could not randomly choose the semester during which the guided method was introduced. This suggests that our results should be interpreted with some caution.

Future research should examine how the quality of learning during experimentation can be affected by one’s approach to the learning task. Students who engage in a task with the aim of learning it (i.e., those with a learning goal orientation) might learn more than those who approach the task “to avoid” disproving their competence. Such differences in learning orientation approaches (see Payne, Youngcourt, & Beaubien, 2007, for a review) can systematically help or hinder the extent to which they can learn from experimentation opportunities. People with a learning goal orientation make plans and set goals when performing a task. Furthermore, they perceive failure as an opportunity to learn, persevere when it happens, and seek feedback. Such an approach is ideal for learning from experimentation.

Experimentation demands goal-setting and planning strategies and can sometimes lead to discovering that one’s hypotheses were incorrect. Those who plan, set goals, perceive failure as an opportunity to learn, persevere when it happens, and seek feedback
are those who would be more likely to learn from activities that teach them how to test hypotheses and develop skills of argumentation. Research in this area will benefit by studies exploring how the effect of structured exploration on argumentative skills is moderated by individual differences in learning orientation.

References


## Description of Dimensions Used to Score Argumentative Term Papers

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descriptive and summarization skills</td>
<td>Articulates hypothesis involving two variables and the relationship between them; provides sufficient and relevant information on evidence for and against hypothesis; is there enough quantity of detail?</td>
</tr>
<tr>
<td>Quality of resources used for definition</td>
<td>Completely references key variables; uses more than one reference per definition to increase the credibility of it; uses academic resources to justify the definitions as they are understood in the field</td>
</tr>
<tr>
<td>Definitional skills</td>
<td>Defines the variables in hypothesis; explains how variables in it are different from closely related variables; describes examples to further the understanding of definitions; comments on the similarities and differences between definitions in the literature; do you understand what the paper is talking about clearly?</td>
</tr>
<tr>
<td>Quality of resources as supporting and countering evidence for hypothesis</td>
<td>Includes expert resources that lend credibility for and against hypothesis; describes resources that have large and reliable samples with credible statistical methods to analyze data</td>
</tr>
<tr>
<td>Quality of resources used for explaining supporting and countering evidence</td>
<td>What are the resources used to explain supporting and countering evidence (i.e., theoretical reviews, meta-analyses, etc.)?</td>
</tr>
<tr>
<td>Quality of resources to reconcile supporting and countering evidence</td>
<td>Uses credible explanations that involve knowledge of universally accepted theories in the field</td>
</tr>
<tr>
<td>Argumentation skills</td>
<td>Is the hypothesis believable? Choose one of the following statements to rate the quality of the hypotheses:</td>
</tr>
<tr>
<td></td>
<td>1. Clear hypothesis, evidence, and explanations with a clear connection (highest quality)</td>
</tr>
<tr>
<td></td>
<td>2. Clear hypothesis, evidence, and explanations without a clear connection or that does not seem to support the hypothesis</td>
</tr>
<tr>
<td></td>
<td>3. Clear hypothesis, evidence, explanations that do not support it or unclear hypothesis with supporting explanations</td>
</tr>
<tr>
<td></td>
<td>4. Clear hypothesis with no evidence or unclear hypothesis with explained evidence that does not support it or unclear hypothesis with supporting evidence but no explanations</td>
</tr>
<tr>
<td></td>
<td>5. Unclear hypothesis with no evidence or explanations (lowest quality)</td>
</tr>
</tbody>
</table>

### Note

Send correspondence to Dr. Phanikiran Radhakrishnan, Department of Management, University of Toronto at Scarborough, Scarborough, ON M1C1A4, Canada; e-mail: phanira@utsc.utoronto.ca
Development and Effects of a Writing and Thinking Course in Psychology
E. Jean Johnson, Albert D. Tuskenis, Glenna L. Howell and Kimberly Jaroszewski

Teaching of Psychology 2011 38: 229
DOI: 10.1177/0098628311421318

The online version of this article can be found at:
http://top.sagepub.com/content/38/4/229

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Society for the Teaching of Psychology

Additional services and information for Teaching of Psychology can be found at:

Email Alerts: http://top.sagepub.com/cgi/alerts
Subscriptions: http://top.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Oct 6, 2011
What is This?
Development and Effects of a Writing and Thinking Course in Psychology

E. Jean Johnson¹, Albert D. Tuskenis¹, Glenna L. Howell¹, and Kimberly Jaroszewski¹

Abstract
The authors developed and assessed a new undergraduate psychology course: Thinking and Writing in Psychology. A description of how the course was developed using the APA learning goals as well as results from an analysis of the course’s effectiveness are offered. The course demonstrated a positive impact on the overall grade point average and thinking and writing skills of students within the course as well as in more advanced courses. Implementing and requiring a thinking and writing course relatively early in the undergraduate psychology major was strongly supported by this study.

Keywords

Are we doing a good job educating our students? How do we know? Historically, there has been a lack of consensus regarding both the components and the assessment of quality in undergraduate psychology programs (Dunn, McCarthy, Baker, Halonen, & Hill, 2007). In 2002, the American Psychological Association (APA) published a set of learning goals and outcomes for the psychology major and an assessment guide for those goals (Halonen et al., 2002). During 2006–2007, our undergraduate psychology program, which is housed at a public university, underwent a review and revision process. We adopted the APA’s learning goals as our program’s goals and integrated them into all program syllabi.

Two APA goals (critical thinking, Goal 3; writing in psychology, Goal 7) were particularly relevant to our university’s interest in a writing-intensive course requirement in each major. We were also influenced by (a) comments from graduate faculty regarding deficiencies in some of our students’ writing and critical thinking skills, (b) the challenge of adopting the APA undergraduate goals throughout our curriculum, (c) national researchers calling for improvement in the writing and thinking skills of psychology majors (e.g., Fallahi, Wood, Austad, & Fallahi, 2006; Ingle, 2007; Reed & Kromrey, 2001; Wade, 1995; Williams, Oliver, Allin, Winn, & Booher, 2003), and (d) discussions in the pedagogy of psychology literature about the current lack of and need for specialized courses in the major (e.g., Dunn, 1994; Goddard, 2003; Lawson, 1999; Penningroth, Despain, & Gray, 2007). We decided that a new thinking- and writing-intensive course at an earlier point in the curriculum would improve the writing and thinking skills needed for success in subsequent more advanced courses. From this confluence of factors, one of our first steps in program revision was to develop a new undergraduate course we titled Thinking and Writing in Psychology.

Thinking and Writing Course Development
As program faculty, we discussed whether to meet the APA goals for thinking and writing through a separate course or across the curriculum. During our discussions, we considered Dunn’s (1994) recommendation regarding an interdisciplinary writing course, suggestions from the Writing Across the Curriculum movement (e.g., Rademacher & Latosi-Sawin, 1995), and ideas for implementing a writing course for psychology majors (Fallahi et al., 2006; Goddard, 2003). Based upon his 1994 review of an interdisciplinary writing course, Dunn recommended that writing experiences for psychology majors include free writing (evaluation free, private, and whatever comes to mind), peer reviews of writing, collaborative writing, and the opportunity for rewriting. Rademacher and Latosi-Sawin (1995) suggested the inclusion of summary writing as well as the use of collaboration and peer review. In their evaluation of the effects of in-class writing instruction in psychology, Fallahi et al. (2006) recommended the inclusion of exercises in grammar, writing style, writing mechanics, and proper use of APA referencing. Goddard (2003) described

¹ Governors State University, University Park, IL, USA

Corresponding Author:
E. Jean Johnson, Division of Psychology and Counseling, Governors State University, 1 University Parkway, University Park, IL 60484-0975
Email: jjohnson@govst.edu
positive effects from an elective writing in psychology course that included exercises in grammar and APA format, collaborative assignments, and peer editing. In addition, she recommended a progressive format (assignments in stages with multiple opportunities for feedback and revision) on primary writing assignments such as a case report, empirical study review and report, conference abstract, and literature review.

In 1999, Lawson described the critical thinking skill type important for psychology majors, which he referred to as “psychological critical thinking” (p. 207). Psychological critical thinking, according to Lawson, involves the ability to evaluate claims in a way that incorporates basic principles of psychological science. In a 2001 study investigating outcomes of teaching critical thinking in a history course, Reed and Kromrey used a model for critical thinking that included elements basic to any reasoning process or task. These included the purpose of the thinking, the question at issue or problem to be solved, fundamental concepts, information, point of view, inferences, assumptions, and implications. They used explicit instruction in critical thinking and instructional activities that included individual, small-group, and total-class components. There were analyses of primary source documents and history problems. In her dissertation research, Ingle (2007) noted that metacognitive self-regulation and elaboration were two robust predictors of critical thinking skills and recommended the use of strategies that would improve those skills, such as problem-based learning strategies.

In addressing how to target Goal 3 (critical thinking) in our curriculum, we agreed that Lawson’s (1999) work on psychological critical thinking as a learning outcome would serve as our base. We also agreed to use additional resources such as Ingle’s (2007) work about critical thinking ability among college students, Reed and Kromrey’s (2001) study on critical thinking in a community college history course, and suggestions from van Gelder (2005) about how to teach critical thinking. Van Gelder recommended the inclusion of deliberate practice exercises in thinking skills, teaching for transfer of skills, directly teaching critical thinking theory, argument mapping exercises, and exercises that challenge students’ existing beliefs. Wade (1995) described a set of writing assignments that were related to the development of critical thinking skills. She gave examples of assignments across various types of psychology courses that would require students to define problems, examine evidence, analyze assumptions and biases, avoid oversimplification, consider alternative explanations, and tolerate uncertainty. Her ideas about the use of writing to develop and assess critical thinking also supported our growing belief in a combined course in both thinking and writing in psychology.

After considerable deliberation, our faculty voted to have APA Goals 3 and 7 integrated both across the curriculum and in a separate specialized course in thinking and writing. We also decided to develop a course that would meet our university outcome requirements for being writing intensive as well as partially satisfying the university’s technology competency requirement.

<table>
<thead>
<tr>
<th>Course Outcome (related APA subgoal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Demonstrate effective critical thinking, including evaluating the quality of information (source, context, and credibility); recognizing common fallacies in thinking; avoiding being swayed by appeals to emotion or authority; synthesizing diverse facts, theories, and observations; and demonstrating an attitude of critical thinking that includes persistence, open-mindedness, tolerance for ambiguity, and intellectual engagement. (3.1)</td>
</tr>
<tr>
<td>2 Practice creative thinking and evaluate one’s own thinking (metacognition). (3.2, 9.2)</td>
</tr>
<tr>
<td>3 Use reasoning to recognize, develop, defend, and criticize arguments and other persuasive appeals. (3.3)</td>
</tr>
<tr>
<td>4 Approach problems effectively. (3.4)</td>
</tr>
<tr>
<td>5 Demonstrate information competence by forming a researchable topic that can be supported by database search strategies and then locating and choosing relevant sources from appropriate media (which may include data and perspectives outside traditional psychology and Western boundaries). (6.1)</td>
</tr>
<tr>
<td>6 Use selected sources after evaluating their suitability based on appropriateness, accuracy, quality, and value of the source; potential bias of the source; and the relative value of primary versus secondary sources, empirical versus nonempirical sources, and peer-reviewed versus non-peer-reviewed sources. (6.1, 7.3)</td>
</tr>
<tr>
<td>7 Use appropriate software to produce understandable reports of the psychological literature. (6.2)</td>
</tr>
<tr>
<td>8 Demonstrate computer skills in basic word processing, database, e-mail, spreadsheet, and data analysis programs. (6.4)</td>
</tr>
<tr>
<td>9 Demonstrate effective writing skills in various formats (e.g., essays, correspondence, technical papers, note taking) and for various purposes (e.g., informing, defending, explaining, persuading, arguing, teaching). (7.1)</td>
</tr>
<tr>
<td>10 Demonstrate effective interpersonal communication skills in articulating ideas thoughtfully and purposefully and by receiving and providing constructive feedback to colleagues in oral and written formats. (7.4, 9.2)</td>
</tr>
</tbody>
</table>

Once our program faculty decided to include a new thinking and writing course in the curriculum, we entered a phase of course development. We began by developing objectives for the course based on relevant APA goals as well as competencies required by our university for all undergraduates in writing-intensive coursework and technology. Each resulting student outcome objective was included in the course syllabus in tandem with its related APA goal and subgoal. These outcomes are listed in Table 1 (with related APA subgoals in parentheses).

After agreeing on the course objectives, we searched for available resources for planning and teaching such a course. We examined the APA annual course frequency listings, but no thinking, writing, or combined course was listed. However, we did find one such course on the Project Syllabus website of the Society for the Teaching of Psychology, Office of Teaching Resources in Psychology (2009). This course, Slattery’s Critical Thinking and Writing in Psychology, served as a model for our new course. We selected some ideas from Slattery’s course.
syllabus that we believed matched the recommended activities from the literature (e.g., Dunn, 1994; Fallahi et al., 2006; Goddard, 2003; Ingle, 2007; Lawson, 1999; Rademacher & Latosinski, 1995; Reed & Kromrey, 2001; van Gelder, 2005; Wade, 1995). We also added assignments and selected texts suggested by various faculty experienced in teaching writing and/or critical thinking.

The new course covered the following topics: structuring the basic psychology paper, critical thinking, retrieving and reviewing library resources and database searches, evaluating sources and evidence, skeptical reading and logical writing, basic research strategies, fundamental grammar, APA style, critiquing research, peer reviewing and editing, preparing drafts, and proofreading the final draft. Assignments for the course preliminarily included assorted written and group critical thinking problems, critical thinking applications (e.g., analyzing a popular press psychology article for fallacies), quizzes (e.g., using the library and conducting searches, critical evaluation of sources, APA format for references and citations), empirical article critiques, and a short literature review. Most of these assignments included the opportunity for rewriting and/or peer review and editing. Class activities were developed to include direct and progressive instruction on writing mechanics and APA style, critical thinking, finding and evaluating evidence, and persuasive and scientific writing style (as opposed to narrative). We began instruction in the new course in the fall term of 2007.

**Assessment Needs**

Several factors led to a general concern about our effectiveness in meeting APA and university standards. These factors included recently published APA quality benchmarks for undergraduate psychology programs (Dunn et al., 2007), the need to link assessment to the psychology learning goals (APA, 2008a), and calls from the APA task force report to strengthen the teaching and learning of undergraduate psychological science (APA, 2008c). The inception of our new curriculum also meant reviewing the current program assessment plan. Effective assessment planning is a major focus of the APA (2008b). As noted previously, there are calls in the literature for improvement in the writing and thinking skills of psychology majors as well as need for specialized courses in the major. However, the literature base is devoid of research on the development and assessment of such courses. There are position papers both regarding the implications of not requiring such rigor and championing the abstract importance of writing courses (e.g., Dunn, 1994; Myers, 2007); however, actual research on the effects of writing courses in psychology is rare. The few publications we did find all included reports of significant positive effects of writing instruction (Fallahi et al., 2006; Goddard, 2003). Specifically, Goddard (2003) found significant effects in student confidence, attitudes, and writing, and Fallahi et al. (2006) documented effects in four writing domains.

The study of critical thinking pedagogy and outcomes yields a similar literature base with a number of position papers (e.g., Reed & Kromrey, 2001; van Gelder, 2005) but only a limited number of empirical research reports (e.g., Lawson, 1999; Penningroth et al., 2007; Williams et al., 2003). Lawson (1999) demonstrated improvement on a measure of critical thinking he created called the Psychological Critical Thinking Exam (PCTE), but it was used as a program outcome measure administered to a limited sample of senior psychology majors (n = 11) and not as a direct result of a course or instruction. More recently, Penningroth et al. (2007) and Williams et al. (2003) used the PCTE to assess psychological critical thinking both as embedded in a developmental psychology course (Williams et al., 2003) and as a separate course designed to improve critical thinking (Penningroth et al., 2007). Students in both studies demonstrated significant gains in critical thinking.

Based on this review, we concluded that we had to plan and develop an original assessment of our new writing and thinking course, and that became the focus of the present study. When we initiated the present study, we were at a transition point in the development of our curriculum between the old and the new curricula; thus, there were two groups of students simultaneously progressing through the program, those completing the new course and those not required to. The group of students not completing the new course, but completing all other courses in the program, functioned as a de facto control group. The purpose of this study was to evaluate the effects of the new thinking and writing course on students’ critical thinking and writing skills (both within the course and in more advanced courses) and general academic performance (as measured by grade point average) and to assess the utility of the course as a prerequisite for more advanced courses within the major.

Research methods courses have been noted to be valuable in achieving the APA learning outcomes (Halonen et al., 2002), meeting quality guidelines (Dunn et al., 2007), and serving as a valuable element in a quality psychology curriculum (APA, 2008c; Dunn et al., 2010). In addition, research methods is one of the most frequent courses in the APA course frequency listings (see Stoloff et al., 2010). Therefore, because they have such strong support in the pedagogical community in psychology and because we believed our research methods courses demand rigorous critical thinking and writing skills, we decided to analyze students’ research methods course performance alone as well as a component of one of four advanced courses.

**Method**

**Participants**

Following human participants review and approval, we began data collection in January 2008 and finished in May 2009. Some 277 undergraduate students in psychology courses and 14 psychology faculty participated. We collected data from two sources: students within the new course and faculty teaching more advanced courses. One of these more advanced courses was Research Methods. Because our university is a nontraditional upper-division school serving junior- and senior-level undergraduates and graduate students, the average age for
student participants of 31 (SD = 9.77) is higher than for traditional students at typical undergraduate institutions. The sample was predominantly female (82%), and the average grade point average was 3.21 on a 4.0 scale (SD = 0.67). In our student database, 187 (67%) psychology students had taken the new Thinking and Writing in Psychology course, and 90 (33%) had not. Faculty participants in our study averaged approximately 15 years of teaching experience in higher education, and more than 70% held terminal degrees.

**Instruments and Procedure**

We used the Ennis–Weir Critical Thinking Essay Test (Ennis & Weir, 1985) as both pretest and posttest for the new course. Students were allowed to decline to participate in the testing. Faculty teaching the courses were not informed about individual student writing or thinking scores. We utilized the published Ennis–Weir rubric for scoring critical thinking to calculate a critical thinking score for each protocol completed. In their 1985 manual, Ennis and Weir claimed good content validity for their instrument and noted high interrater reliabilities of .86 and .82. After additional validation research, Ennis (2005) stated there was “strong content-related support and moderate empirical support for a claim for its [Ennis–Weir’s] situational validity” (p. 3).

In addition to scoring for critical thinking using the protocol provided in the manual, we also required a measure of writing skills. The decision was made to use the Ennis–Weir essays to create two scores, writing and critical thinking. One of the grant team faculty members with doctoral degrees in both psychology and language arts created a rubric to also score writing from the Ennis–Weir protocols. The writing rubric assessed seven areas of writing ability: (a) content knowledge, (b) thesis, (c) organization/evaluation, (d) clarity/coherence/vocabulary, (e) grammar/mechanics/usage, (f) major sentence-level errors, and (g) fluency. Writing was evaluated for each of these on a 0–3 scale, with 0 indicating that the skill addressed was inadequate, unprofessional, or lacked understanding; 3 indicating that the skill assessed was fluent and professional (virtually error-free) and demonstrated thorough mastery; and 1 and 2 indicating skill levels in between these poles. Interrater reliability calculations on the rubric’s results established significant concordance (ICC = .84, p = .001) and reliability (α = .95) for the four of us as the faculty readers and scorers. The faculty readers were blind as to whether the essay being scored was a pre- or posttest.

In all sections of the new course during the grant period (fall 2007 through spring 2009), students completed pre- and posttests scored for critical thinking and writing. Research faculty reviewed student records for gender, age, and grade point averages. We monitored course completers’ grade point averages and grades in more advanced courses 1–2 semesters after the course.

We surveyed program faculty teaching four different higher level courses for which most students may have completed the new thinking and writing course regarding student performance (grade for the higher level course), apparent critical thinking skills, and demonstrated writing skills. Of these four courses, the new course is listed as a prerequisite for three. These courses were Principles of Learning and Behavior (new course; not a prerequisite), Professional Standards (ethics), Psychological Issues and Values, and Research Methods. We formatted the surveys as class lists on which the faculty reported the student’s final grade and assigned two scores (0–10, where 0 represented incapable and 10 represented exceptional) indicating their judgment of each student’s thinking and writing abilities. We did not inform the faculty teaching the four advanced courses whether any student on the list had completed the thinking and writing course. We then searched the university record system and determined whether each student had completed the thinking and writing course, verified each student’s grades for all pertinent courses, and determined demographics (age and gender) and previous and current grade point averages. We assessed effects within the course on thinking and writing skills, differences between new course completers and noncompleters in thinking and writing skills and grade point averages, and effects on grades and thinking and writing skills in research methods courses. We also evaluated the viability of the new course as a prerequisite for the more advanced courses in question. There were no significant differences on any measure on the basis of age or gender.

**Results**

**Effects Within the Course on Thinking and Writing Skills**

Repeated measures t tests were used to compare pretest and posttest scores on thinking and writing from the Ennis–Weir testing. We had 70 student participants complete both administrations. The average pretest thinking score was 5.93 (SD = 5.83) and posttest thinking scores averaged 7.61 (SD = 5.96). The average pretest writing score was 7.30 (SD = 3.47), and posttest writing scores averaged 8.54 (SD = 3.52). The paired samples t test indicated significant improvement within the course from pretest to posttest in both thinking, t(69) = 2.68, p = .009, d = .285, and writing, t(69) = 3.00, p = .004, d = .355.

**Differences Between Completers and Noncompleters**

**Thinking and writing scores.** Faculty scores for thinking and writing skills in higher level courses were collected. Students may have been scored by a faculty member for one to four courses. If a student was scored in more than one advanced course, we averaged his or her scores. Average thinking and writing skill scores were compared between completers and noncompleters of the new course. Independent measures statistics were calculated. Of the 196 students scored by the faculty, 109 students had completed the new course and 87 students had not. Thinking and writing skills were scored higher for those students who had completed the new course. For writing skills, completers had an average score of 7.66 (SD = 1.74) and
noncompleters averaged 6.59 (SD = 2.20), t(194) = 3.79, p = .001, d = .54.

Because of the presence of a significant Levene statistic for the thinking skill scorings, we used a more conservative statistic with a reduced degree of freedom. For thinking skills, completers had an average score of 7.80 (SD = 1.65) and noncompleters averaged 6.82 (SD = 2.24), t(154) = 3.41, p = .001, d = .499.

Grade point averages. From the faculty ratings of thinking and writing in advanced courses, we determined viable grade point averages for 210 students, 184 of whom had completed the new course and 26 of whom had not. We excluded students who were enrolled in the more advanced courses simultaneously with the new course. The average grade point for completers was 3.25 (SD = 0.64) and for noncompleters was 2.92 (SD = 0.81), t(208) = 2.44, p = .016, d = .45.

Differences Within Research Methods Courses
Grades for the Research Methods courses. We obtained the final grades for 135 students completing a section of Research Methods. Of those 135 students, 80 had completed the new writing course prior to taking Research Methods and 55 had not. The average Research Methods grade for new course completers was 3.19 (SD = 0.98) and for noncompleters was 2.67 (SD = 1.56), t(130) = 2.78, p = .006, d = .49.

Writing skills. We obtained faculty ratings for writing skills for 112 students completing a section of Research Methods. Of those 112 students, 58 had completed the new writing course prior to taking Research Methods and 54 had not. The average writing skill score for new course completers was 7.66 (SD = 1.93) and for noncompleters was 6.46 (SD = 2.34), t(110) = 2.95, p = .004, d = .56.

Thinking skills. We obtained faculty ratings for thinking skills for 113 students completing a section of Research Methods. Of those 113 students, 59 had completed the new writing course prior to taking Research Methods and 54 had not. The average thinking skill score for new course completers was 7.68 (SD = 2.06) and for noncompleters was 6.76 (SD = 2.41), t(111) = 2.18, p = .031, d = .42.

Viability as a Prerequisite for Higher Level Courses
Grades. We compared student grades in the new course with their grade(s) in any more advanced course(s) completed afterward. For each of the four higher level courses examined, a student’s performance (grade) in the new course was found to be a significant or marginally significant predictor of performance in the advanced course (p values = .001 to .060). In all cases, the direction of relationship was positive, meaning that higher grades in the writing course were associated with higher grades in the more advanced course. Although grades in different psychology courses would intuitively be significantly and positively correlated, the focus here was on how well the new course functioned as a prerequisite. One way to discern this was the comparison of completers to noncompleters previously discussed. Another was to ascertain whether the new (prerequisite) course was a significant predictor and the degree to which it accounted for variation in the more advanced course.

For the Principles of Learning and Behavior course, the writing course was only a marginally significant predictor, F(1, 83) = 3.64, p = .060, r² = .04. This is the only advanced course of the four for which the new writing course is not currently a prerequisite. For the Professional Standards course, the new writing course was a significant predictor of the student’s grade, F(1, 58) = 12.10, p = .001, r² = .17. This course is the second writing-intensive course in the program (the new writing course is the first of the sequence). For the Psychological Issues and Values course, the new writing course was a significant predictor of the student’s grade, F(1, 72) = 7.35, p = .008, r² = .09. For the Research Methods course, the new writing course was a prerequisite and a significant predictor of the student’s grade, F(1, 43) = 11.38, p = .002, r² = .21. A review of the R-squared values indicated between 4% and 21% of the variation in the grades for the advanced course was accounted for by the thinking and writing course. The strongest relationships and variation accounted for were for the Professional Standards and Research Methods.

Timing of course and grades. We assessed the timing of students’ taking the thinking and writing course in relation to their grades in advanced courses. Whether students took the new course before, concurrently with, or after the advanced course was a significant factor for grades in some of the advanced courses. Because of some small cell sizes and the presence of some significant Levene statistics, we calculated Kruskal–Wallis analysis of variance (ANOVA) statistics. Grades were higher for those students taking the advanced course after or at the same time they took the thinking and writing course. For two courses, the Learning and Behavior course (new thinking and writing course not a prerequisite) and Research Methods (new thinking and writing course a prerequisite), there were significant differences in grades for students taking the advanced course after or concurrently with the new thinking and writing course rather than before it. Grades for the Professional Standards and Issues and Values courses were higher when the student took the course concurrently or after the thinking and writing course, but not significantly so. Results of the analysis are summarized in Table 2.

Timing of the course and thinking and writing skills. We investigated the timing of each student’s taking the thinking and writing course as a factor with regard to thinking and writing skills in more advanced courses (as scored by the faculty). Because of some small cell sizes and the presence of some significant Levene statistics, we used Kruskal–Wallis ANOVAs. When students took a higher level course (i.e., before the thinking and writing course, concurrently, or after the thinking and writing course) was a significant factor in only one case: Writing skill scores for students in the research methods courses were
significant higher after taking the new thinking and writing course, $\chi^2 (2, n = 62) = 6.74, p = .039$. In all other cases, though not significantly so, there was general improvement in both thinking and writing scores for each higher level course when the student took the higher level course after or concurrently with the new thinking and writing course.

**Discussion**

Following calls for research and development in the writing and thinking skills of psychology majors, we developed a new undergraduate psychology course. The development and assessment of this new thinking and writing course was the focus of the present work. According to Halpern (2004), “[T]here is no single indicator of educational effectiveness . . . no single number can capture the multifaceted nature of an effective college education or the cognitive growth of an adult student. . . . [And] successful assessment programs are owned by the faculty” (pp. 16-17). In the Assessment Cyberguide, APA (2008a, 2008b) task force members noted that assessment is most beneficial when it is embedded within existing coursework and part of an assessment plan that focuses on only a few goals (subsets that may change and rotate) on an annual basis. The work being done for this course and program assessment was multifaceted but limited to two goals. We intend this assessment to evolve into an integral component of an ongoing program assessment plan, embedded within the thinking and writing course (as well as outside the course) and conducted by the faculty, thereby meeting all of the requisite challenges posed for curricular assessment by APA.

Our new course appears to have significant effects on students’ writing and thinking skills and grade point averages. These results mirror numerous previous findings for writing (e.g., Fallahi et al., 2006; Goddard, 2003; Rademacher & Latosi-Sawin, 1995) as well as critical thinking (e.g., Williams et al., 2003). One of the challenges that Goddard (2003) raised concerned the “ideal placement of the course in the psychology curriculum” (p. 28). Our study appears to contribute one answer to this question. Although not significant in all cases, there was general improvement in both thinking and writing skill scores as well as in grades for all higher level courses based on whether and when the students took the new thinking and writing course. Although academic maturation as well as increasingly more developmental curricula may suggest an inherent progression in thinking and writing skills of psychology undergraduates, the comparison of completers to noncompleters suggests the new course augments this process.

Similar to Williams et al. (2003), who found that critical thinking was a moderate predictor of course performance, we found that our thinking and writing course also appears to have significant predictive validity for grades in three higher level courses for which it is currently a prerequisite ($R^2 = .04$ to $.21$). Although not significant, it also has some predictive ability for thinking and writing skills in general. Both the faculty involved in developing the course and other program faculty members believe the new course should remain a prerequisite for the three higher level courses.

Our study’s results and the conclusions drawn must be considered in the context of possible limitations. First, although overall sample size was adequate ($N = 277$), statistical analyses based on reduced cell sizes because of some significant Levene tests (homogeneity of variance) and multiple comparisons required nonparametric measures for some analyses and may have reduced some statistical power. Second, the students in this sample were not provided any incentive (beyond the faculty’s good will) to participate. We administered the pretest at the start of the term when students were new to the course and more concerned with gaining faculty approval. Unfortunately, we had to administer the posttests, near the end of the term when students were more likely to be encumbered with numerous assignments and less concerned with faculty approval. Anecdotal comments from faculty indicated students seemed to spend less time on the posttests than on the pretests. Thus, we assume that the scores for the pretest were more indicative of a “good effort.” Our results may, therefore, actually underestimate improvements in writing and thinking skills. Future researchers may wish to provide an incentive for students to produce a “good effort” at both pre- and posttesting.

Third, the faculty teaching the advanced courses were asked to provide scores (0–10) indicative of what they considered to be the students’ thinking and writing skills. Although our “definitions” of those scores were provided to the faculty in descriptive form, the scores were both subjective in nature (i.e., individual faculty determined) and obtained at least 2 weeks after the completion of the advanced course or courses. An objective instrument, one delivered at the start of the advanced course, may be more valid in assessing thinking and writing skills as a direct result of the thinking and writing course.

Finally, support for the validity of the Ennis–Weir Critical Thinking Essay Test (Ennis & Weir, 1985) appears only

---

**Table 2. Advanced Course Grades Related to When New Thinking and Writing Course Taken**

<table>
<thead>
<tr>
<th>Course</th>
<th>Mean Grade Rank</th>
<th>n</th>
<th>$\chi^2$</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning and Behavior</td>
<td>Before 61.19</td>
<td>72</td>
<td>6.17</td>
<td>2</td>
<td>.046</td>
</tr>
<tr>
<td></td>
<td>During 77.29</td>
<td>42</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>After 75.64</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Professional Standards</td>
<td>Before 45.90</td>
<td>40</td>
<td>1.98</td>
<td>2</td>
<td>.372</td>
</tr>
<tr>
<td></td>
<td>During 54.22</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>After 47.74</td>
<td>34</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Issues and Values</td>
<td>Before 36.57</td>
<td>27</td>
<td>3.25</td>
<td>2</td>
<td>.197</td>
</tr>
<tr>
<td></td>
<td>During 39.63</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>After 45.87</td>
<td>31</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Research Methods</td>
<td>Before 35.06</td>
<td>42</td>
<td>9.27</td>
<td>2</td>
<td>.010</td>
</tr>
<tr>
<td></td>
<td>During 37.24</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>After 51.78</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Before = took advanced course before new course; During = took concurrently; After = took advanced course after new course.
moderate, and for assessing critical thinking specifically within the domain of psychology, it may be particularly questionable. We used the Ennis–Weir instead of the PCTE (Lawson, 1999) because of the Ennis–Weir’s essay format. The PCTE uses a short-answer format that would make scoring for writing doubly difficult to accomplish. However, our creation of a rubric to also score the Ennis–Weir for writing requires validation. To address these measurement concerns, the development of an instrument easier to use and score as well as more dedicated to measuring the writing (i.e., APA style) as well as thinking skills relevant for psychology (such as those measured by the PCTE) is now our goal.

Based on this ongoing analysis, we have implemented a number of ideas. In early sections of the new thinking and writing course, grades for the course averaged 3.36. Faculty involved in developing and teaching the course concluded there should be more emphasis on evaluation of students’ writing and thinking in the courses and less on group work (e.g., workshops of each other’s writings, group critical thinking exercises) and class participation. The average for grades has since declined to 2.84. From our sample, faculty feedback, student evaluations of class participation. The average for grades has since declined to 2.84. From our sample, faculty feedback, student evaluations of instruction, and comments from the Goddard (2003) study, we also concluded that more focus on APA-style writing (and actual practice) is needed. In the course sections taught most recently, more writing, peer review, and attention to APA style were added, resulting in a significant decrease in course grades, hypothetically because of the more stringent requirements, $F(10, 157) = 4.61, p = .001$, $\eta^2 = .23$. At this time, we expect to continue with these more stringent assignments. In addition, the graduate program faculty are recommending that the college consider the use of this thinking and writing course for graduate student remediation.

To review, two of the APA’s undergraduate psychology learning goals include critical thinking and writing in psychology. Psychology faculty desiring to meet these goals in their programs should be encouraged and supported in the development of writing and thinking in psychology courses. Such efforts will need to address issues such as deficiencies in students’ writing and critical thinking skills and the general need for improvement in the writing and thinking skills of psychology majors. Courses in the major at earlier points in undergraduate programs should also be encouraged and supported in the development of writing and thinking in psychology courses. Our development and assessment of the course indicates support for direct and progressive (developmental) instruction in topics such as the basic psychology paper, critical thinking, library resources and database searches, evaluating sources and evidence, skeptical reading and logical writing, basic research strategies, basic grammar, APA style, critiquing research, peer reviewing and editing, and preparing drafts. Assignments for such courses should include a number of assorted written and group critical thinking problems, critical thinking applications (e.g., analyzing a popular press psychology article for fallacies), quizzes (e.g., using the library and conducting searches, critical evaluation of sources, APA format for references and citations), empirical article critiques, and a short literature review. In addition, most of these assignments should include the opportunity for rewriting and/or peer review and editing.

Our development and evaluation of the effects of a new thinking and writing course in psychology has been overwhelmingly positive. The course has demonstrated a positive impact on overall grade point average and writing and thinking skills. Our new thinking and writing course has viability as a prerequisite for more advanced courses. The development, implementation, and requirement of a thinking and writing course relatively early in the undergraduate psychology major appear to have relatively strong support. The development of an instrument to accurately measure the thinking and writing skills of psychology and social science majors is also supported and encouraged.

Authors’ Note
This article is an expansion and elaboration of a research poster presented at the 117th annual meeting of the American Psychological Association in August 2009, Toronto, Ontario, Canada.

Declaration of Conflicting interests
The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The authors received some support for this research in the form of a University Research Grant from Governors State University.

References


Practice Makes Perfect: Improving Students' Skills in Understanding and Avoiding Plagiarism With a Themed Methods Course
Sarah Estow, Eva K. Lawrence and Kathrynn A. Adams
Teaching of Psychology 2011 38: 255
DOI: 10.1177/0098628311421323

The online version of this article can be found at:
http://top.sagepub.com/content/38/4/255

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Society for the Teaching of Psychology

Additional services and information for Teaching of Psychology can be found at:

Email Alerts: http://top.sagepub.com/cgi/alerts
Subscriptions: http://top.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Oct 6, 2011
What is This?
Faculty Forum

Practice Makes Perfect: Improving Students’ Skills in Understanding and Avoiding Plagiarism With a Themed Methods Course

Sarah Estow1, Eva K. Lawrence1, and Kathrynn A. Adams1

Abstract
To address the issue of plagiarism, students in two undergraduate Research Methods and Analysis courses conducted, analyzed, and wrote up original research on the topic of plagiarism. Students in an otherwise identical course completed the same assignments but examined a different research topic. At the start and end of the semester, all students (n = 44) completed a homework assignment assessing plagiarism knowledge and paraphrasing skills. Students in the plagiarism-themed courses showed improvement in both knowledge and skills, and the strategies they suggested for avoiding plagiarism became more sophisticated as did the reasons for avoiding plagiarism. The control group did not show the same improvements. Results suggest repeated hands-on exposure to the topic of plagiarism improves plagiarism avoidance and understanding.

Keywords
academic integrity, plagiarism reduction, paraphrasing

Educators are currently "at war" with academically dishonest acts, including plagiarism (Leask, 2006). By many accounts, plagiarism is a pervasive problem that is only becoming more prevalent on college campuses (Lim & See, 2001; Macdonald & Carroll, 2006; Pickard, 2006; Roig, 1997; Roig & Caso, 2005; Schuetze, 2004). Although most students would agree that it is dishonest to copy an entire paper and submit it as one’s own work, many students do not recognize that citation and proper paraphrasing are key components of researching with integrity (Roig, 1997). For example, students in Singapore considered paraphrasing without a citation as the least serious type of academic cheating with 90% admitting that they had done so (Lim & See, 2001). One might interpret the lack of citation as a sign that students knowingly flout the rules of good academic work, but some evidence suggests that citation problems emanate from a lack of understanding (Belter & du Pré, 2009; Culwin, 2006; Landau, Druen, & Arcuri, 2002). Moreover, academic codes of conduct may lack sufficient detail and clarity about academic dishonesty (Leask, 2006; Roig, 2001). Faculty may disagree about whether specific behaviors (e.g., self-plagiarism) constitute plagiarism (K. K. Bennett et al., 2011; Culwin, 2006; Lim & See, 2001; Schuetze, 2004). However, research shows that merely lecturing to students or providing resources explaining what academic dishonesty is and how to avoid it are not as effective as hands-on experience (Culwin, 2006; Macdonald & Carroll, 2006; Schuetze, 2004). For example, even a brief homework assignment on plagiarism can reduce its occurrence (Belter & du Pré, 2009; Schuetze, 2004).

Using a more immersive approach, we developed and implemented a semester-long plagiarism theme with multiple assignments in a methods and statistics course. Using a common theme in a research methods course provides course continuity, a larger, shared participant pool, and quality control in the use of primary research articles (Marek, Christopher, & Walker, 2004). We expected that using plagiarism as a common theme in our methods and statistics course would change the way that students viewed plagiarism as well as improve their skills in avoiding plagiarism.

1 Guilford College, Greensboro, North Carolina, USA

Corresponding Author:
Sarah Estow, Psychology Department, Guilford College, 5800 W. Friendly Ave., Greensboro, NC 27410
Email: sestow@guilford.edu
Materials and Design

Students completed homework assignments at the beginning (pretest) and end (posttest) of both sections to assess knowledge of plagiarism. On both the pretest and posttest, students listed strategies for avoiding plagiarism and reasons why it was important to do so. Students also paraphrased a sentence about plagiarism and identified which of three versions listed below the original sentence illustrated plagiarism. For the pretest, the sentence to be paraphrased was taken from Schuetze (2004); for the posttest, it was taken from R. Bennett (2005). Appendix A illustrates the pre- and posttest materials.

Design and Procedure

The quasi-experimental study used a 2 (Time of Test: Pretest and Posttest) × 2 (Course Topic: Plagiarism or Nonplagiarism) mixed design. The syllabi included a statement that all work should be done according to the principles in the college-wide academic honor code and referred students to the student handbook for further information. Students in both conditions completed the identical pretest homework early in each semester. After the students turned in the pretest assignment, we discussed the importance of plagiarism avoidance in both conditions. In the plagiarism condition, as we covered various study designs and statistics during the semester, a variety of assignments were related to the plagiarism theme (see Table 1). In the nonplagiarism condition, the course structure and required assignments were the same. However, the assignments related to the effects of gender and ethnicity on teaching evaluations rather than to plagiarism. Students in both conditions completed the identical posttest homework at the end of the semester.

Results

We counted the number of sentences correctly identified as plagiarism on both homework assignments. Results of a 2 (Time) × 2 (Topic) mixed ANOVA revealed a significant interaction, $F(1, 42) = 9.19, p = .004, \eta^2 = .18$. Students in the plagiarism group had higher variance during pretest, although the mean scores between the classes on pretest were not significantly different, $t(38.51) = -1.97, p = .056$. At posttest, the plagiarism-themed class had higher scores than the control class, $t(42) = 2.42, p = .020$ (see Table 2).

One of the researchers rated the students’ own paraphrasing of the source sentence on a 4-point scale from 1 (direct copying of a significant portion of the original without quotation marks) to 4 (good paraphrasing). The $2 \times 2$ mixed ANOVA results

Table 1. Course Assignments for Plagiarism Themed Course

<table>
<thead>
<tr>
<th>Student Activity Related to Plagiarism</th>
<th>Details of Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analyzed article (Schuetze, 2004)</td>
<td>Identified rationale, hypothesis, IV, DV, operational definition of plagiarism knowledge, findings, and future directions.</td>
</tr>
<tr>
<td>Conducted interviews</td>
<td>Questioned four participants regarding definition of plagiarism, confusion about plagiarism, reasons for plagiarism, and ways to reduce plagiarism.</td>
</tr>
<tr>
<td>Wrote APA-style report about interviews</td>
<td>Required citation to Schuetze (2004) and minimum of one additional source.</td>
</tr>
<tr>
<td>Developed survey and collected data</td>
<td>Surveyed four participants about their knowledge of, and attitudes toward, intentional and unintentional plagiarism.</td>
</tr>
<tr>
<td>Formulated descriptive hypotheses</td>
<td>One hypothesis pertained to the entire sample, and one compared subgroups of the sample.</td>
</tr>
<tr>
<td>Conducted descriptive analyses using SPSS</td>
<td>Selected appropriate variables and descriptive statistics based on hypotheses.</td>
</tr>
<tr>
<td>Wrote APA-style research report</td>
<td>Used scenario to manipulate type of plagiarism (serious or minor) and intent (intentional or unintentional); DVs were perceived mood of student and professor and severity of punishment, with each student questioning four participants.</td>
</tr>
<tr>
<td>Conducted experimental research</td>
<td>Required information from at least two primary research articles.</td>
</tr>
<tr>
<td>Analyzed data using SPSS and prepared experimental report</td>
<td>Selected one experimental IV, analyzed results using an independent-samples t test, wrote research report incorporating at least three primary research articles.</td>
</tr>
<tr>
<td>Analyzed data using SPSS and prepared second report</td>
<td>Selected two IVs (one possibly non-experimental), analyzed results using a two-way ANOVA, and wrote a research report incorporating at least four primary research articles.</td>
</tr>
</tbody>
</table>

Note: The control group did the same assignments but focused on a different research topic.

Method

Participants

Forty-four undergraduate psychology majors in research methods courses at a small Southeastern liberal arts college completed the study by submitting both pre- and posttest data, 27 in the plagiarism theme condition (Spring 2007), and 17 in the nonplagiarism theme condition (Spring 2008). The cumulative GPAs for students in the plagiarism theme condition ($M = 2.85, SD = .53$) and in the nonplagiarism theme condition ($M = 3.09, SD = .43$) were similar, $t(42) = 1.61, p = .12, d = .43$. An additional 18 students provided only pretest data primarily because of withdrawals. The withdrawal rate was not atypical for this particular course.
Results indicate that giving students hands-on experience and repeated exposure to the topic of plagiarism improves their ability to identify faulty paraphrasing, enhances their own paraphrasing, and deepens their understanding of why one should avoid plagiarism. These students were primarily juniors and seniors who had taken at least two other psychology courses before methods and statistics. Thus, it is somewhat disheartening that initial paraphrasing and plagiarism identification scores were so low despite the institutional honor code and strict plagiarism policy. For many of our students, paraphrasing and appropriate citation appeared to be relatively new skills consistent with Macdonald and Carroll’s (2006) research on plagiarism. Regardless, our findings are encouraging; immersing students in the topic of plagiarism had clear benefits and improved their skills significantly. Both groups received explicit plagiarism education at the beginning of the semester, and both improved their ability to identify plagiarism. However, the control group’s quality of paraphrasing and number of strategies listed to avoid plagiarism actually decreased slightly from beginning to end of the course when plagiarism was not an explicit focus (see Table 2). These findings underscore the importance of a continued focus on plagiarism throughout the semester.

Moreover, the immersion approach appeared to deepen how students conceptualized plagiarism and its reduction. Because the first article the students analyzed (Schuetze, 2004) focused on plagiarism and students explored intentional versus unintentional plagiarism, it is not surprising that, by the end of the semester, many more students noted the importance of education as a way to avoid plagiarism. However, we never explicitly told students that education or practice was the “correct” strategy to avoid plagiarism, nor were they aware that they would be completing the homework assignment at the end of the semester. Thus, the fact that 56% of students mentioned educational strategies on the posttest may represent internalization of important concepts. We would expect this number to be even higher had we addressed this topic explicitly. Posttest responses also revealed expanded appreciation of the ethical issues surrounding plagiarism, another topic that we did not discuss explicitly.

One limitation of this study is that we required students to paraphrase only a single sentence at a time. However, it seems reasonable to consider that a plagiarism-themed course would also enhance students’ ability to paraphrase more comprehensive text. An additional limitation is that we only had a single rater coding the open-ended responses. Despite these limitations, both of which may be addressed by future research, we recommend that instructors consider incorporating a plagiarism theme in methods courses. Such a theme advances multiple objectives: enhancing knowledge of plagiarism, illustrating how researchers build research programs, and highlighting multiple perspectives on the same topic.

<table>
<thead>
<tr>
<th>Outcome:</th>
<th>Plagiarism Theme M (SD)</th>
<th>Control M (SD)</th>
<th>Plagiarism Theme M (SD)</th>
<th>Control M (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plagiarism identification</td>
<td>1.52 (0.98)</td>
<td>1.94 (0.43)</td>
<td>2.59 (0.75)</td>
<td>2.06 (0.66)</td>
</tr>
<tr>
<td>Quality of paraphrasing</td>
<td>2.54 (0.86)</td>
<td>3.12 (0.78)</td>
<td>3.23 (0.86)</td>
<td>2.82 (0.81)</td>
</tr>
<tr>
<td>Number of strategies</td>
<td>2.44 (0.93)</td>
<td>2.69 (0.60)</td>
<td>2.63 (1.18)</td>
<td>2.00 (0.73)</td>
</tr>
</tbody>
</table>

Discussion

Table 2. Assessment of Student Learning in Plagiarism-Themed Course Compared to a Control Course

McNemar’s nonparametric tests were used to examine the specific strategies and reasons listed to avoid plagiarism. In the plagiarism-themed course, three (11%) listed education as a strategy on the first homework, whereas 15 (56%) listed this on the second homework, \( \chi^2(1) = 8.64, p = .002 \). On the other hand, only two (13%) in the control group listed educational strategies on both pre- and posttest, \( \chi^2(1) = .50, p = 1.00 \). On both the first and second assignment, the majority of students in the plagiarism-themed course (66% and 56%) and the control (69% and 88%) indicated that avoiding punishment was a major reason not to plagiarize. There was no significant change from pre- to posttest for the plagiarism-themed course, \( \chi^2(1) = .11, p = .58 \), or the control, \( \chi^2(1) = 2.01, p = .25 \). However, students in the plagiarism-themed course were more likely to add that plagiarism should be avoided because it is unethical (i.e., it hurts the original author, the college, or the field of psychology) from pretest \( (n = 17, 63\%) \) to posttest \( (n = 26, 97\%) \), \( \chi^2(1) = 7.11, p = .004 \). The control class did not demonstrate this increase (pre-: \( n = 7, 43\% \); post-: \( n = 4, 25\% \), \( \chi^2(1) = .57, p = .45 \).
Appendix A

Sentences Used in the Pretest and Posttest to Assess Paraphrasing and Plagiarism Identification Skills

**Pretest**
Sentence students paraphrased:
Increased student confidence in their ability to avoid plagiarism would hypothetically result in an inaccurate perception that they are fully knowledgeable about the complexities involved in proper citations in scientific papers (Schuetze, 2004, p. 259).

Sentences students assessed for plagiarism:
Increased student confidence in their ability to avoid plagiarism would hypothetically result in an inaccurate perception that they are fully knowledgeable about the complexities involved in proper citations in scientific papers.

Increased student confidence in their ability to avoid plagiarism would hypothetically result in an inaccurate perception that they are fully knowledgeable about the complexities involved in proper citations in scientific papers (Schuetze, 2004).

One danger of increasing students’ confidence in their ability to avoid plagiarism is that this overconfidence could leave them unaware that they do not understand the complexities of proper citation (Schuetze, 2004).

**Posttest**
There is a need for university staff to address forcefully the issue of academic integrity during introductory programs and to explain clearly and sympathetically the objective need for honesty in academic life (R. Bennett, 2005, p. 156).

University staff should be very clear and sympathetic about the importance of academic integrity, and they should do this early in a student’s academic career.

There is a need for staff of universities and colleges to forcefully address the issue of academic honesty during beginning programs and to explain clearly and sympathetically the objective desire for integrity in college life (R. Bennett, 2005).

Declaration of Conflicting Interests
The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The authors received no financial support for the research, authorship, and/or publication of this article.

References


A Reliable and Valid Weighted Scoring Instrument for Use in Grading APA-Style Empirical Research

Report

Kathleen Puglisi Greenberg

Teaching of Psychology 2012 39: 17
DOI: 10.1177/0098628311430643

The online version of this article can be found at:
http://top.sagepub.com/content/39/1/17
A Reliable and Valid Weighted Scoring Instrument for Use in Grading APA-Style Empirical Research Report

Kathleen Puglisi Greenberg

Abstract
The scoring instrument described in this article is based on a deconstruction of the seven sections of an American Psychological Association (APA)-style empirical research report into a set of learning outcomes divided into content-, expression-, and format-related categories. A double-weighting scheme used to score the report yields a final grade that reflects the relative importance of outcomes in each category and the differential contribution of each section of the report to the report as a whole. The scores produced by the instrument are reliable between and within raters and significantly correlated with students’ cumulative grade point averages. The author hopes the instrument can provide a useful framework for scoring any set of learning outcomes an instructor defines as the essential elements of an APA-style research report.

Keywords
grading, scoring, scoring rubrics

One of the first topics addressed in many introductory research methods texts is the notion of objectivity and the integral role it plays in the process of scientific inquiry (e.g., Beins, 2009; Christensen, Johnson, & Turner, 2011; Goodwin, 2008; Neuman, 2011). However, the importance of objectivity extends beyond the realm of science and into the domain of the research methods classroom itself when those of us who teach this course take on the task of grading the American Psychological Association (APA)-style research reports that we often require our students to write. In light of this interconnectedness between what we tell our students about the importance of objectivity and what we actually do when we grade their papers, it may be said that those of us who teach research methods have a special obligation to “practice what we preach” and grade our students’ reports as fairly and objectively as possible.

One way to approach the issue of objectivity in grading is through the use of grading rubrics (Moskal, 2000; Peat, 2006). However, despite their widespread use in both K-12 and higher education settings, it is just recently that we have begun to see efforts to create rubrics for the specific purpose of grading students’ empirical research reports. Gottfried, Vosmik, and Johnson (2009) described the process they used to develop a rubric of this kind. Despite their very thoughtful approach, they did not find high levels of interrater reliability in the scores the rubric produced. Nonetheless, they argued persuasively that rubric development should continue in light of the many benefits that rubrics provide (see Malini, 2010, for a review of rubric use in higher education).

More recently, there have been two reports of rubrics developed for grading APA-style research papers. Stellmack, Konheim-Kalkstein, Manor, Massey, and Schmitz (2009) created a rubric for grading APA-style introductions in which eight dimensions are rated on a 4-point scale of achievement. Thaler, Kazemi, and Huscher (2009) designed one for grading an entire research report, using a 6-point Likert scale to gauge the achievement of 10 learning outcomes based directly on the guidelines provided in the fifth edition of the Publication Manual of the American Psychological Association (APA, 2001). Measures of convergent validity were strong for both rubrics; however, only the Thaler et al. rubric was found to have significant interrater reliability (as measured by the correlation between the scores of different raters).

Comparison of the two rubrics suggests that one possible source of the discrepancy in findings is a difference in the complexity of the rating tasks. The rating scale used by Stellmack et al. (2009) consists of four qualitatively different levels of achievement. However, despite their widespread use in both K-12 and higher education settings, it is just recently that we have begun to see efforts to create rubrics for the specific purpose of grading students’ empirical research reports. Gottfried, Vosmik, and Johnson (2009) described the process they used to develop a rubric of this kind. Despite their very thoughtful approach, they did not find high levels of interrater reliability in the scores the rubric produced. Nonetheless, they argued persuasively that rubric development should continue in light of the many benefits that rubrics provide (see Malini, 2010, for a review of rubric use in higher education).
best describes the extent to which an outcome has been achieved. When an achievement category is defined by one or two criteria, the task is fairly straightforward. For example, the “sources” learning outcome in the Stellmack et al. rubric is given the highest achievement rating based on (a) the number of sources referenced and (b) the extent to which these sources are relevant to the topic and cited in the paper. However, when an achievement category is based on several criteria—as is often the case for higher levels of proficiency—the rater’s task is complicated by the need to acquire a holistic understanding of the set of criteria that defines each of the four levels of achievement, and also by the need to decide how many criteria within a category must be met for that level of achievement to be considered fair and appropriate. In the Thaler et al. (2009) rubric, the levels of the rating scale vary (essentially) on a single continuum of “sufficiency.” This use of a single continuum gives the scale quantitative properties that allow the rater to make what seems, on judgment, like a simpler decision about the extent to which the outcome has been achieved. With a simpler task, differences between raters based on proficiency with the rubric would be minimized, as would any differences associated with the interpretation of the scoring categories, particularly when a category is defined by multiple criteria that may have been only partially met.

If there is some merit to this conceptualization of the two types of rubrics, it suggests that the utility of a rubric as an objective and reliable means of scoring written work may be limited by the degree to which it is designed with largely qualitative or quantitative scaling properties. A rubric designed with scoring categories defined by an amalgam of learning criteria is inherently a qualitative measuring instrument; as such, it would be expected to yield scores that differ more between raters than would be the case with a rubric in which outcome achievement is rated on a single dimension.

In addition to issues with weak reliability, the “qualitative” type of rubric does not provide students with feedback as to which criteria within a scoring category they have met; nor, of course, does it provide information about the extent to which each of those criteria has been mastered. It is also more difficult for students to use this type of rubric proscriptively because the scoring categories (or at least the category that defines the highest level of achievement) must be deconstructed in order for students to be able to identify the criteria that will be used to evaluate their work.

Consideration of these limitations of qualitative rubrics as scoring devices led to the development of the scoring instrument described here. It is referred to as a scoring instrument, and not as a rubric, because each learning criterion is rated separately and on a single dimension of “achievement.” The main benefits of the instrument as a way of grading APA-style research reports are that (a) it makes explicit the learning criteria upon which the grade is based; (b) it delineates those criteria in a way that makes it easy for students to use them as a guide when preparing their reports; (c) it provides specific feedback regarding the extent to which each learning criterion has been achieved; (d) it yields subscores for each section of the report so students can see which sections were done well and which may need some work; (e) it produces scores that reflect the relative importance of content-, expression-, and format-related outcomes; (f) it results in a final grade that reflects the differential contribution of each section of the report to the report as a whole; and (g) it standardizes the scoring process so that the subscores and final grade it produces are reliable both between and within raters, whether the scoring is done by an experienced instructor or an advanced undergraduate student.

The purpose of this article is to describe the instrument and present data demonstrating significant interrater and intrarater reliability and significant criterion validity as well. I hope the description can serve as a framework that can be adapted to fit whatever set of learning outcomes an instructor defines as the essential elements of an APA-style research report.

The Scoring Instrument

The version of the instrument described here consists of a set of 60 learning outcomes that collectively represent what I expect my students to include or show evidence of in the research reports they write. The outcomes associated with each section of the report (i.e., title page through references section) fall into three superordinate categories—Content, Expression, and Format—based on whether they respectively pertain to what was said, how it was said, or whether the text was formatted properly. (See Table 1 for a model of the introduction section of the instrument.) Each outcome is rated on a 4-point scale from 0 (absent/not at all) to 3 (completely) based on the extent to which the work provides evidence that the outcome has been achieved. The ratings can be written on a two-page score sheet given to students, or entered directly into an Excel spreadsheet used to calculate the scores. After the outcomes are rated, the ratings within each category are summed and converted to a percentage of possible points earned. This percentage is then weighted to reflect the relative importance of each category. For example, one might assign weight of 0.5 for content-related outcomes, 0.3 for those that are expression-related, and 0.2 for those associated with formatting. The weighted outcomes are calculated separately for each section of the report, yielding a set of Section Subscores that provide students with a measure of their performance for each section. The final scoring is done by weighting each Subscore to reflect the impact it is to have on the final grade (with the body of the report presumably being given the most weight). Copies of the instrument, the two-page score sheet, and the Excel file are available at http://goo.gl/ckb5n.

Data Collection and Analysis

For this study, 20 papers were selected randomly from a set of 45 first-and-only drafts submitted by students in two research methods classes at a small public college in suburban New York. Each class was composed primarily of female students
and students in their sophomore or junior year of study. Both were taught by the same instructor, and this instructor served as Rater 1 (R1). Another instructor who taught different sections of the same research methods courses as did R1 served as the second rater (R2). R1 and R2 scored the papers in July 2009. In May 2010, the papers were scored by a third rater (R3) who, at the time, was a graduating senior in the psychology department. The second scoring of the papers by R1 took place in December 2010, roughly 18 months after the original scoring. (For ease of communication, the first and second scoring of the papers by R1 will respectively be referred to as R1 Time-1 and R1 Time-2.) All three raters were familiar with the scoring instrument (R2 used it for two semesters, and R3 used it in an introductory and an advanced research methods course), so very little discussion about how to use it took place other than to briefly review the definitions of each outcome. A person not associated with the research recorded the names of the students whose papers were scored and blackened them out with a permanent marker before distributing the papers for scoring.

### Interrater Reliability

Table 2 shows the levels of interrater reliability (as measured with Pearson’s r) for the Final Score, Section Subscores, and mean Category Scores for each pair of raters (R1-R2, R1-R3, and R2-R3). There is a high level of reliability overall, as indicated by average correlations of r(18) = .84 for R1-R2, r(18) = .65 for R1-R3, and r(18) = .74 for R2-R3. All but three individual coefficients are statistically significant (p < .05), and most are higher than .70. Two of the three that did not reach statistical significance were those for the Expression category scores for the two instructor-student pairs of raters (R1-R3 and R2-R3). The lack of statistical significance in this particular area suggests that the scoring of expression-related outcomes may be more subject to individual differences in judgment than those in other categories and that instructors and students may have different standards for what constitutes effective writing. Nonetheless, the high level of interrater reliability between each of the three pairs of raters indicates that the instrument promotes a high level of consistency in grading and that it does so regardless of whether the papers are scored by someone with expertise in the subject matter or someone whose exposure to it is relatively limited.

### Table 1. Model of the Introduction Section of the Scoring Instrument

III. Introduction → [Goal: provide empirical context and explain study’s purpose]

<table>
<thead>
<tr>
<th>A. Content [points earned: ____ /out of 12 → percentage: ____%] × 50% =</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. General orientation</td>
</tr>
<tr>
<td>introduction of topic; definition of variables being studied</td>
</tr>
<tr>
<td>2. Empirical context</td>
</tr>
<tr>
<td>well-researched, focused literature review using primary source information</td>
</tr>
<tr>
<td>3. Purpose</td>
</tr>
<tr>
<td>accurate and clear statement of study’s purpose</td>
</tr>
<tr>
<td>4. Hypothesis</td>
</tr>
<tr>
<td>accurate statement of hypothesis(es) being tested, with obvious rationale</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B. Expression [points earned: ____ /out of 9 → percentage: ____%] × 30% =</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Organization of ideas</td>
</tr>
<tr>
<td>ideas/thoughts flow logically from paragraph to paragraph; paragraphs have a topic sentence and supporting details</td>
</tr>
<tr>
<td>2. Mechanics/clarity</td>
</tr>
<tr>
<td>rules of grammar are followed; punctuation is correct and appropriate; words are spelled correctly</td>
</tr>
<tr>
<td>3. Voice</td>
</tr>
<tr>
<td>no slang; no informal phrases (e.g., ended up); no connection with the reader (e.g., “you”), etc.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>C. Formatting [points earned: ____ /out of 6 → percentage: ____%] × 20% =</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Title</td>
</tr>
<tr>
<td>study title, written in title case; no bold</td>
</tr>
<tr>
<td>2. Citations</td>
</tr>
<tr>
<td>written in Modified Harvard (name, date) citation format</td>
</tr>
</tbody>
</table>

Note: Outcome achievement rating scale ranges from 0 (absent/not at all) to 3 (completely).
Intrarater Reliability

An intrarater reliability analysis for the Final Score, Section Subscores, and mean Category Scores (again, as measured with Pearson’s $r$) showed that the scores generated by the instrument are highly consistent from Time-1 to Time-2. All coefficients are statistically significant ($p < .01$) and, with the exception of the Expression category scores, greater than or equal to $r(18) = .84$.

Criterion Validity

Criterion validity was assessed by measuring the relationship between the Final Scores and the cumulative grade point averages (GPAs) of 10 students whose GPA information could be obtained. Although this clearly is a very limited sample, the correlation between the two measures is significant ($p < .05$) for the two instructor raters, and marginally significant ($p = .084$) in the case of the student rater. Specifically, for R Time-1, $r(8) = .75$; for R1 Time-2, $r(8) = .76$; for R2, $r(8) = .80$; and for R3, $r(8) = .57$. Although the small sample makes it impossible to say with confidence that the scores generated by the instrument are predictive of overall academic performance, the findings nonetheless are consistent with this conclusion.

Subscores and Final Grades

Although the inter- and intrarater reliability data indicate that the instrument promotes relative consistency in scoring both across and within raters, the question remains as to whether the Final Scores generated by the instrument are of the same magnitude regardless of who has scored the paper. The results of a MANOVA conducted on the Final Scores and the mean Category Scores show that the Time-2 scores for R1 are not significantly different from the scores for R2 or R3 and, furthermore, that the scores for R2 and R3—with the exception of the mean Expression score—are not significantly different from each other. Specifically, the MANOVA provides evidence for an overall mean difference among the raters, $F(8, 108) = 4.00$, $p < .01$, that is traceable (with follow-up univariate analyses and Tukey post hoc comparisons) to a sole difference between the mean Expression Category Scores for Raters 2 and 3, $F(2, 57) = 4.30$, $p < .05$. All the remaining scores—including the Final Scores—are not significantly different across raters (see Table 3). Thus, it does appear that the instrument yields scores that are neither statistically nor practically different regardless of whether the scoring is done by the course instructor, another instructor, or an upper-division student who has used the instrument in class.

One final question is whether the scores generated by the instrument differ, on average, when the papers are graded by the same instructor at two different points in time. Table 4 shows the Final Score and mean Category Scores for the first rater at Time-1 and at Time-2. A series of paired $t$ tests revealed that the scores at Time-1 are significantly lower than the scores at Time-2 (all $ps < .01$) by about 15 points (translating into a difference of about 1.5 letter grades on a standard 100% scale). Thus, although the instrument produces scores that are reliable over time (as indicated by the intrarater reliability data), it clearly does not serve to eliminate all sources of subjectivity in the grading process.

Two possible sources of such subjectivity are an instructor’s expectations and standards for defining acceptable performance. The nature of the discrepancy between the scores at Time-1 and Time-2 suggests that R1 may have used a stricter standard when scoring the papers for the first time than when scoring them 18 months later. Given that R1 was the instructor for the course from which the papers had been selected, and that the papers were first scored not long after they had been submitted, it is possible that the stricter standard reflects a tendency to have high expectations when papers are initially graded because one so readily remembers what students were asked to do and what was done to help them do it. (How many times have we found ourselves perplexed and even a bit frustrated when a student provides an incorrect answer to a question on a topic that we know for certain was well-covered in class?) Thus, it follows that when there is no such recollection (as presumably would be the case 18 months after the papers were originally scored), expectations “return” to a level that is consistent with those of other instructors who lack this personal point of reference. Admittedly, this line of reasoning is highly speculative (and circular); however, it is supported by the data in Table 3, showing no differences between the

---

**Table 3.** Final Scores and Mean Category Scores by Rater (R1, R2, R3)

<table>
<thead>
<tr>
<th>Category</th>
<th>R1 Time-1</th>
<th>R2</th>
<th>R3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>Final Score</td>
<td>77.9</td>
<td>13.0</td>
<td>76.0</td>
</tr>
<tr>
<td>Mean Category Scores</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Content</td>
<td>78.7</td>
<td>13.1</td>
<td>79.8</td>
</tr>
<tr>
<td>Expression</td>
<td>80.2</td>
<td>15.9</td>
<td>74.1</td>
</tr>
<tr>
<td>Formatting</td>
<td>86.5</td>
<td>11.0</td>
<td>84.5</td>
</tr>
</tbody>
</table>

Note: Univariate test results for each dependent measure are as follows: Final Score: $F(2, 57) = .790, p > .05$; mean Content score: $F(2, 57) = .261, p > .05$; mean Expression score: $F(2, 57) = 4.298, p < .05$; mean Formatting score: $F(2, 57) = .277, p > .05$.

*Significantly different ($p < .05$) from the mean for R2 as determined by Tukey post hoc comparisons; all remaining comparisons revealed no significant differences.
Mean Category Scores

scoring devices that vary along a qualitative-quantitative
to the quantitative end, and the Thaler et al. rubric at some point
in between. Given that quantitative measures are more reliable
than qualitative measures, it is perhaps not surprising to see
gradations in reliability across the three measurement devices,
with a lack of reliability reported by Stellmack et al., and strong
reliability reported by Thaler et al., and moderately strong reliability reported by Thaler et al., and strong
reliability in the case of the current instrument.

If there is validity to this conceptualization of rubrics as
scoring devices that vary along a qualitative-quantitative
measurement continuum, it may not be appropriate to think
of holistic and analytical rubrics as representing mutually
exclusive categories. In the literature, rubrics are often charac-
terized as being either holistic or analytical depending on
whether they respectively yield a single score or multiple
scores (e.g., Kan, 2007; Mertler, 2001). However, it could be
argued that what differentiates one from the other is not
whether the work is evaluated as a whole or in parts, but the
extent to which individual learning criteria are combined to
form more complex, or higher-order, learning outcomes. In a
holistic rubric, there is one such outcome, defined on the basis
of the entire set of established learning criteria. However, an
analytical rubric can have any number of outcomes, from just
two (if the set is described by only two dimensions) to as many
outcomes as there are in the entire set (at which point, it may
not be considered a rubric at all). Thus, in the characterization
of a rubric, it may not be a matter of deciding whether it is
holistic or analytical, but rather, if it is not holistic (i.e., quali-
itative), then how analytic (quantitative) is it?

What is relevant about this discussion is the suggestion that
achieving a level of objectivity in grading students’ APA-style
research reports with the use of a rubric (or scoring instrument)
may require a highly analytic—even molecular—approach in
which the assessment criteria/learning outcomes are minimally
dimensionalized, if at all. In addition to bringing a level of
objectivity to the report grading process, a scoring instrument
of this kind has the additional advantage of providing students
with both detailed guidance and specific feedback to help them
produce a high-quality report the first time they write one and
an even better one the next time around. Tutors and instructors
also benefit from the detailed feedback, as it serves to pinpoint
areas of weakness and make the remediation process poten-
tially more efficient and effective. Specific advantages offered
by the current instrument include the “dimensional” feedback
students receive in the form of the Section Subscores and the
scores within the Content, Expression, and Format achieve-
ment categories, along with the assurance that the grade they
earn is a fair and appropriate reflection of the relative impor-
tance of these three types of outcomes and the degree to which
each section of the report contributes to the quality of the report
as a whole.

Although the instrument was designed to assess the achieve-
ment of outcomes comprising this type of report, if the ability
to write one is a key learning outcome of the research methods

### Summary and Conclusions

The weighted scoring instrument described in this article was
developed in an effort to promote objectivity in the scoring of
APA-style research reports. An analysis of 20 papers scored
by three raters—the course instructor, an instructor not
associated with the course, and an upper-division psychology
major—showed that the instrument has a high degree of inter-
and intrarater reliability, and that the scores it yields are signif-
ically correlated with students’ cumulative GPA. Moreover,
an intrarater analysis of final scores suggests that the instru-
ment does produce scores that are not significantly different
across raters, (presumably) as long as the expectations and
standards of those raters are aligned.

A comparison of the current scoring instrument with the
rubrics designed by Stellmack et al. (2009) and Thaler et al.
(2009) suggests that the reliability of a scoring rubric may, in
part, be a function of whether the rating scale it uses is essen-
tially qualitative or quantitative in nature. When outcome
achievement is scored categorically on the basis of multiple cri-
teria, the rating task is more qualitative than quantitative; when
the learning criteria are scored separately, as is the case with
the current instrument, the task is more quantitative. In this
sense, the Stellmack et al. rubric, the Thaler et al. rubric, and
the current scoring instrument represent three places along a
continuum of measurement, with the Stellmack et al. rubric
closest to the qualitative end, the current instrument closest
to the quantitative end, and the Thaler et al. rubric at some point
in between. Given that quantitative measures are more reliable
than qualitative measures, it is perhaps not surprising to see
gradations in reliability across the three measurement devices,
with a lack of reliability reported by Stellmack et al., moderately
strong reliability reported by Thaler et al., and strong
reliability in the case of the current instrument.

If there is validity to this conceptualization of rubrics as
scoring devices that vary along a qualitative-quantitative

### Table 4. Final Scores and Mean Category Scores for Rater 1 (R1) at Time-1 and Time-2

<table>
<thead>
<tr>
<th></th>
<th>R1 Time-1</th>
<th>R1 Time-2</th>
<th>( t )</th>
<th>( p ) Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Final Score</strong></td>
<td>63.2</td>
<td>77.9</td>
<td>-8.808</td>
<td>.001</td>
</tr>
<tr>
<td><strong>Mean Category Scores</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Content</td>
<td>67.3</td>
<td>78.7</td>
<td>-7.397</td>
<td>.001</td>
</tr>
<tr>
<td>Expression</td>
<td>59.5</td>
<td>80.2</td>
<td>-4.949</td>
<td>.001</td>
</tr>
<tr>
<td>Formatting</td>
<td>82.8</td>
<td>86.5</td>
<td>-3.764</td>
<td>.001</td>
</tr>
</tbody>
</table>

Time-2 scores of R1 and the scores obtained by the other
two raters. Nonetheless, it will be important in future
research on the instrument to explore the role of instructor
expectations and standards as possible sources of score
variability, even when attempts are made (as with the cur-
rent instrument) to adopt a standardized approach to gener-
ating them.
course, the scores produced by the instrument also can be used to provide a holistic assessment of the degree to which this broader outcome has been achieved. This can be done by establishing categories of achievement each defined by a range of scores and calculating the percentage of students whose grades fall within each category. These data can then be used as the basis for a comparative assessment across instructors, within instructors, and across introductory and advanced research methods courses with an APA-style report course requirement.

Limitations

In addition to limitations associated with the small sample size, there are limitations specific to the instrument as well. Clearly, the content of an APA-style research report is indicative of a student’s ability to demonstrate a number of higher-order skills, including those related to critical thinking, quantitative reasoning, and effective writing. As such, it may not be possible to reduce these complex skills into a number of discrete elements without losing the essence of what these skills represent (Sadler & Good, 2006). On the other hand, from a strictly behavioral point of view, it could be argued that it is possible to teach complex behavior by identifying and shaping its individual components, much like it is possible to teach someone to play the piano by first having them learn to play individual notes. Nonetheless, to the extent that a high-quality APA-style report, like a piano concerto, is more than the sum of its parts, it is acknowledged that the instrument may be better suited for scoring fundamental outcomes than for scoring outcomes that reflect higher-order, and more abstract, cognitive processes.

Another potential limitation concerns the issue of how time-consuming and possibly tedious it is to rate a large number of outcomes. In actuality, with a bit of practice (and presumably less than what is required to become proficient in the use of a qualitatively scored rubric), the scoring process becomes fairly automatic, partly because the criteria to be scored are precisely the elements the instructor is looking for and expects to see in the report. For example, in the current instrument, there are 15 outcomes associated with the formatting of the references section. They reflect every feature (however minor) of a properly formatted reference, from the use of hanging indents and double-spacing to the proper formatting of the authors’ names. (Even the use of an ampersand between the authors who might wish to use a version of the instrument in their courses to further investigate its utility and pedagogical value. Special thanks are due to Nancy Bray and Tressa Cincotta for scoring the papers, B. Runi Mukherji for her valuable comments on an earlier version of the scoring instrument, and Coleman Paul and William S. Altman for their thoughtful comments on an earlier draft of the article.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

Notes

2. A copy of the Thaler, Kazemi, and Huscher (2009) rubric can be found at http://docs.google.com/Doc?id=df0b863n_0dw8dm32g.
3. The instrument described here is based on the guidelines provided in the fifth edition of the Publication Manual of the American Psychological Association (American Psychological Association,
The instrument I currently use has been modified to reflect the revisions in the sixth edition of the *Publication Manual* (American Psychological Association, 2010) and changes in my thinking about the importance and operationalization of certain outcomes.

4. The mean was computed by averaging the scores within each achievement category across all sections of the report.

**References**


ties, and the value of activities that are more obviously applicable to student instruction.

The quandary that emerged from the teaching–research–service model, wherein students, faculty, and administrators place different value on professors’ out-of-class activities, is in contrast to the model that emerged from the factor analysis. The second model suggests a number of helpful solutions. To achieve a compromise among the needs of students, faculty, and administrators, professors can engage in research that involves students, is conducted at a level that is engaging to students, and addresses not only the educational experience of students but also the importance of college and community service. Professors can also prompt classroom discussions on the value of student guidance that takes place outside the classroom. Faculty would also benefit by communicating to administrators the need to acknowledge the energy expended in individual contact with students outside of the classroom, which these student data suggest is the most important aspect of the undergraduate education.

References


Note

Send correspondence to Christopher A. Bjornsen, Department of Psychology, Longwood College, 201 High Street, Farmville, VA 23909; e-mail: cbjornse@longwood.lwc.edu.

Attending a Professional Conference: A Hands-On Seminar Course

Judith R. Levine
State University of New York, Farmingdale

In this article, I describe a course built around participation in a teaching of psychology conference. It presents the rationale, goals, and requirements of the course as well as the benefits to students.

Since 1987 my department has run an annual teaching of psychology conference. The conference takes place over a 3-day period, and attendees are together for formal sessions as well as meals, breaks, and informal gatherings.

Since 1992 students have had the opportunity to receive three credits for attending this conference by enrolling in an academic course. The only prerequisite is a previous psychology course. This academic course provides undergraduates a unique opportunity to become involved in a professional experience.

Course Rationale

State University of New York, Farmingdale is a specialized college offering primarily 2- and 4-year degrees in the technologies and other applied fields. A psychology major is not offered. Students interested in psychology would likely enroll for a 2-year degree in Liberal Arts and Sciences and complete only two or three psychology courses.

Others have described the benefits to students of attending regional and undergraduate research conferences (e.g., Lipton, 1986). Tryon (1985) emphasized the opportunity to learn about current research findings at regional conferences. Carsrud, Palladino, Tanke, Aubrecht, and Huber (1984) stressed the value of undergraduate research conferences to students for developing competence in research methodologies and statistical techniques. Anderson and Rosenfeld (1983), however, believed there was too much emphasis on research at undergraduate conferences, often resulting in boredom. Because my students’ backgrounds in psychology are limited, both regional conferences and undergraduate research conferences, with their emphases on statistics and methodology, would be inappropriate. However, my colleagues and I believe a variety of students can benefit from exposure to a professional conference. A conference on teaching psychology is an ideal vehicle for this exposure because even nonmajors are well acquainted with teaching and learning.

Course Goals and Objectives

I expect students to acquire knowledge in content areas of psychology and relate this new information to material that they have learned in other psychology courses. The specific knowledge each student acquires depends on the presentations that student chooses to attend.

A second goal is for students to demonstrate communication skills. (In Spring 1999 this course was designated as fulfilling part of SUNY Farmingdale’s communication skills requirement.) I expect them to read the initial proposals submitted by presenters, listen to the presentations critically, and convey in their papers and class discussions what they have learned using a clear and coherent organizational structure and relevant supporting information.

My department developed this course, in part, with the objective of changing students’ awareness of and attitudes toward learning and the educational process. Specifically, students should develop a critical awareness of what goes on in the classroom in the hope that by understanding teaching, they will become better learners. In addition, by observing the process of professional development, students will dis-
cover that learning in general and job training in particular are life-long processes. Students always have new information to learn and new techniques to master.

For students who are considering careers in psychology or teaching (after additional education), the conference provides an opportunity for reality testing to determine whether they identify with these professional groups. Thus, the interpersonal interactions between students and faculty form an important part of this course. The conference also affords them something of an inside look into the job of college professor (Zechmeister & Reich, 1994).

Course Requirements

Written Assignments

Students receive copies of the initial proposals for each scheduled presentation. They must read each proposal, determine which presentations they will attend, and write a summary of each one. Each summary includes the topic, method of presentation, whether it was on the conference theme, and why it was of interest. The summaries are worth 30% of the final grade.

A final paper (10 to 15 pages), worth 60% of the final grade, is due at the postconference meeting (see the next section). I expect this paper to include (a) a critical summary and analysis of each presentation attended, the conference as a whole, and the course; (b) a statement explaining how the student’s views on the teaching–learning process were influenced by this conference; and (c) a description of how conference participation influenced the student’s educational or career goals.

Meetings and Attendance

Students must fulfill several attendance requirements. First, they attend three preconference meetings (5 weeks, 3 weeks, and 1 week before the conference) and one postconference meeting (4 weeks after the conference) on campus.

At the first meeting, I distribute course outlines, requirements for the first written assignment, and the proposals. I explain the rationale, goals, requirements, and mechanics of the course. In explaining the assignment, I provide details about the nature and purpose of professional conferences in general and this one in particular. These details include discussing the conference theme and describing the different types of sessions as well as discussing the purpose of proposals submitted by the presenters.

At the second meeting, I distribute hotel registration forms and travel directions. Although students must register and pay for their own hotel rooms and provide their own transportation, they are automatically registered for the conference at no charge. The rest of the meeting is devoted to discussing the proposals, and I collect the summaries for grading.

At the third meeting I return the graded summaries. Then I review rules of conduct, including such things as the permissibility of tape recording sessions and proper dress. Finally, I explain the requirements for the final paper.

These meetings allow students and me to get acquainted. We will spend 3 long, hectic days together, and it is important to develop a sense of community. At the postconference meeting students present a summary of their final papers and turn them in for grading.

Another attendance requirement is that each student be present for the entire conference and attend five sessions in addition to three keynote addresses. To facilitate professional identification, I require my students to mingle with a variety of professors during meals, and I monitor the situation to ensure that they do.

Attendance at a course meeting each night after dinner is also mandatory. Students discuss their opinions and insights regarding the sessions they attended and the conference as a whole. Attendance and participation count for 10% of the final grade.

Evaluation

I use class discussions to evaluate students’ oral communication skills. These discussions also serve as a vehicle through which students formulate and clarify ideas for their final papers. I use the final paper as the primary mechanism to evaluate the degree to which the course goals are met. The quality of writing and the issues that students must address in the paper speak specifically to the course goals. It is informative to note the sessions that students have attended and analyzed. As a group, they have preferred sessions dealing with learning styles and strategies, gender issues and courses, and technology in the classroom. However, students have also attended sessions on statistics, color-mixing phenomena, social comparison theory, social conformity, and ego defense mechanisms, among others. Although the emphasis of most presentations is how to teach these topics, students do learn about the topics themselves. In all cases, students are learning about learning. Students successfully communicated what they learned and distinguished this information from their opinions.

Students’ comments in their papers illustrate that they gained a greater understanding of the educational process, professional development, and career options in psychology. Representative comments follow:

The conference gave me a totally different view of my professors and their integrity concerning the teaching practice. I want to be a lawyer, but after attending the conference I’m considering a doctorate in psychology as well as my degree in law.

The whole reason for [the conference] was to try to find new ways to motivate and guide students. I know that the insights I’ve gained will definitely be of great advantage in my future endeavors [as a teacher].

Eight of the 11 students who took the course in Spring 1995 and 9 of the 17 who took it in Spring 1996 completed a brief anonymous course evaluation form. In general, they believed that they learned more (71%) from this course than
An Exercise in Unreliability

J. Ronald Gentile
Counseling and Educational Psychology
State University of New York at Buffalo

In this article, I describe a classroom exercise that demonstrates the unreliability of essay scoring. Designed as a 30 to 45-min exercise for a class in tests and measurement, teams of students invent keys or rubrics and score other students’ brief essays. Unknown to the various examiner teams, the examinees have submitted the same exam for scoring; thus it is the examiner teams being tested, not the examinees’ essays. Once the ruse is exposed and students see the wide variety of scores and scoring keys that they assigned to the same essay, discussion about issues of reliability and validity follows naturally and meaningfully.

One of the challenges in tests and measurement courses, particularly those aimed at prospective teachers, is to convince students of the difficulty of scoring essays reliably and validly. One can lecture about the problem, discuss the importance of pre-established criteria and rubrics, and even evoke memories of students’ own perceived horror stories. Still, students believe that they somehow will be immune to the problem or that their errors will be insignificant.

Sax (1980) demonstrated the problem quite well. He obtained 3 to 4 sentence essays from three sixth graders on whether adding two numbers must always give a smaller answer than when the same numbers are multiplied. He then gave these essays to 38 student teachers to score on a 0 to 10 scale, with no further instruction except to disregard spelling, grammar, and other aspects of style in their scoring. Sax reported, “The raters assigned all possible scores from 0 to 10 to each paper” (p. 138). Furthermore, the children’s teacher originally assigned scores of 6, 4, and 4 to the papers—and 1 day later rescored them 5, 8, and 2, respectively. Sax’s results replicated those of Starch and Elliott (1912, 1913a, 1913b), who may have been the first investigators to bring unreliability of essay scoring under scientific scrutiny.

Other noteworthy classroom exercises on scoring or interpreting tests are in Cohen (1996), especially “The Cohen Chicken Soup Essay” (pp. 12–13) designed to demonstrate the importance of cultural expectations in testing, and Forgan (1973), who demonstrated the feelings associated with labeling by administering a group intelligence test to teachers in his class and assigning each the same IQ of 87.

I developed a classroom “exercise in unreliability” that is both effective and dramatic. It takes 30 to 45 min and can be used with large classes. I begin by announcing that we are going to have an in-class practice exercise on scoring essays, and then I follow these steps:

1. Divide students into groups of 3 to 5, widely separated from one another. Have each group select a volunteer to leave the room for a special assignment.
2. Send the volunteers to another room, taking pen and paper, to await your instructions.
3. After the volunteers have left, display a course-relevant question (see Table 1A for a sample) on the overhead projector and, with your best malicious look or evil grin, say that you created that question as a surprise test for the volunteer examinees. This disclosure usually elicits sighs of relief from the group members (“I’m glad it wasn’t me,” etc.). Announce that their job as an examining committee is to create a key or rubric for scoring the essay that the examinees will turn in to them in about 15 min. Emphasize that they will not only have to score the essay, but that as good teachers they will have to explain the scoring system and give examinees feedback on their essays. If there are questions about the purpose of the assignment, honestly answer that you wish for them to see the wide variety of scoring methods that will likely arise from a variety of independent examiners.
reduce content validity (see Cronbach, 1970; Wesman, 1971). Items that are better discriminators may test more difficult concepts or may contain more difficult wording (e.g., double negatives, longer stems, or fine distinctions between choices; see Cronbach, 1970). In the case of wording difficulty, reading ability or logical reasoning ability may influence a student’s answer as much as or more than the student’s content knowledge and thus represents what Cronbach called an irrelevant difficulty. Thus, instructors who go to great lengths to write course exams that contain items with high discrimination indexes, or who restrict their exams to test bank items with high indexes, may inadvertently narrow their pool of items to one with a higher incidence of such difficulties. One end result is that higher grades may be awarded to students as much for their verbal and analytical skills as for their content knowledge. At an extreme, such outcomes could contribute to what some intelligence quotient critics have called an academic elitism (Lemann, 1999), in which it is largely those who “test well” (see Stanley, 1971) or who have high abilities in narrow domains (e.g., as tested by the Scholastic Aptitude Tests) who make it to the better universities or graduate schools. Assigning course grades does require some discrimination among students, but the range of difficulty in concepts and the differing levels of students’ studying and recall seem sufficient to ensure a reasonable range of grades.

Conclusions

In sum, near-zero or negative discrimination indexes can identify test items that are ambiguous or low in content validity, and these items deserve removal. In the absence of these factors, however, these items should probably be maintained. Otherwise, there are potential risks to content validity and to the subset of students who tend to answer such items correctly. In any case, the statistical assumption in this exclusion strategy, that a single construct is being measured, seems questionable for course exams (although factor analytic approaches might partially address this issue). Given in part these issues, I suggest a focus on content validity in the construction of course exams and encourage careful consideration of the potential causes of low discrimination indexes.

References


Use of In-Class Lab Groups to Enrich Independent Research Projects

Cynthia L. S. Pury
Clemson University

In this article, I suggest that the independent project required by most methodology courses can be profitably combined with a lab group model. Such in-class lab groups offer students the chance to explore an area of individual interest, while providing a forum for immediate peer feedback. In addition, students benefit from in-depth exposure to their group members’ projects.

Independent research projects are frequently assigned in research methodology courses (e.g., Chamberlain, 1986; Yoder, 1979). Such projects enhance student interest and allow instructors to evaluate students’ application of course material. The typical structure of independent projects, including midsemester design and individual consultation, demands much instructor time and can lead to student apprehension (Chamberlain, 1986).

Group projects offer an alternative (e.g., Chamberlain, 1986; Dunn, 1996). They teach teamwork and allow students to create more complex projects by division of labor. However, assigning individual grades for group projects is difficult, and social loafing can become problematic (Meyers, 1997).

Small, independent project lab groups can achieve many of the goals of independent and group projects, while avoiding many of their disadvantages. I have used such groups successfully for both introductory methodology and
Groups are comprised of four to six students. Each student develops his or her own research project. During each stage of the project, students discuss their plans and achievements with fellow group members and the instructor. Groups meet with the instructor during part of the regularly assigned class time (30 to 60 min per group per week). The small enrollment of my courses (12 to 20 students) permits me to meet with each group every time; for courses with larger enrollment, many of the same techniques could be implemented by having lab groups meet with teaching assistants or even independently.

I assign students to groups following the first class meeting based primarily on similar research interests. When I know students from previous classes, I use a modified jigsaw approach as well by placing more- and less-able students in the same group. Each group meeting follows the same general format: Students bring in a written assignment, present it to the group, and get feedback. Students can incorporate group feedback into their assignment before it is graded.

For the methodology course, students complete the following assignments:

1. Students discuss research ideas with their group and begin to develop a topic. Starting independent projects at the beginning of the semester underscores the important point that research methods do not exist in a dull vacuum; rather, they are tools designed to help scientists answer interesting questions. I ask students to apply newly presented designs and methods to their project during lecture. In this way, the independent projects serve as the problems for an extended problem method (McBurney, 1995).
2. Students outline at least three relevant studies from primary source journal articles.
3. Students write a research question.
4. Students suggest ways to measure constructs needed to answer their question.
5. Students submit a hypothesis.
6. Students submit a draft of their introduction section.
7. Students submit a draft of their methods section. Feedback from group members is especially helpful, as the group provides knowledgeable pilot participants. Members who have similar projects may decide to gather data together.
8. Students submit a revised introduction and methods sections.
9. Students complete an IRB form (see Kallgren & Tauber, 1996). Students submit this form to a different lab group in the course, which serves as a student-run IRB.1 In this way, students apply ethical principles to studies other than their own.
10. Students gather data and plan statistical analyses. The first session after data collection usually provides excellent support for novice experimenters, who frequently are disheartened at experimental no shows and surprised by unexpected participant behavior.
11. Students complete data entry. Analyses are conducted during lab, with group members observing.
12. Students submit a draft of their results section.
13. Students outline their discussion section.
14. Students bring in any questions or revised section drafts. Most students are pleasantly surprised to discover that their final project is all but completed.
15. Students turn in their final paper and present their project to the entire class.

For the psychometric course, the basic structure of the group is similar to the introductory methodology course, although students develop their own reliable, valid test of a construct rather than test a hypothesis. Because of this emphasis on test development, the psychometrics groups focus more on data collection and statistical procedures and less on writing than the methodology groups.

In-class lab groups provide research collaboration; peer pressure to motivate work; and frequent feedback, which are benefits of regular research group meetings as described by Horner, Stetter, and McCann (1998). In addition, weekly assignments teach students to tackle large projects in small steps. These benefits come without the difficulty of assessing each member's role in a joint project. Rather, the group serves a more supportive function, and the instructor's role becomes more facilitative (McKeachie, Lin, Moffett, & Daugherty, 1978). The group provides multiple viewpoints on students’ projects and in-depth exposure to additional applications of course material. Discussion of the assignments before submission also simplifies weekly feedback for the instructor.

The major difficulty in conducting lab groups is encouraging group feedback. Meyers (1997) commented on a similar problem in collaborative learning efforts and suggested that students need to believe that their contributions are important. I have found it extremely useful to borrow from the problem-method approach (McBurney, 1995) and remind students that there is no one correct answer as to how to design and run a research project. In fact, their feedback to their classmates may be something I had not considered. Following each student assignment presentation, I also ask each member individually to comment on what they have just heard. Finally, I make it clear that 10% of each student's grade is based on my evaluation of the quality of their contribution to lab group; thus, providing an incentive for participation (McBurney, 1995).

Evaluations of both courses have been positive.2 In the Introduction to Methodology course, the mean rating of the course was 4.5 on a 5-point scale ranging from 1 (not at all, indicating low satisfaction) to 5 (very much, indicating high satisfaction), with a modal response of 5 and no student giving any rating below 3. In the Psychometric course, the

---

1Clemson University allows instructors to approve projects that are conducted as part of a class requirement, with the instructors accountable to the Internal Review Board.

2Due to institutional changes, student evaluation forms were slightly different for each course. Thus, only the overall mean for all questions of course quality are included.
mean rating of the course was 4.7, with a modal response of 5 and no student giving a rating below 4. In the free-response portion of the evaluations, 6 of 15 introductory methodology students listed the lab group as one of the strengths of the course; none listed it as a weakness. Results from the psychometrics course were even more positive, with 7 of 10 participants listing the lab group as a strength of the course or a particularly helpful teaching method. None listed it as a weakness. Selective comments from both courses included “Good professor–student interaction, especially in lab, offers a chance for deeper understanding of material”, “I really liked the way lab was set up. It really helped with the independent project”; and “Taking on the semester-long project in small steps was very helpful and a good strategy.”

In summary, an independent project lab group can provide students with a supportive, helpful environment to decrease the stress associated with tackling the independent research project (e.g., Chamberlain, 1986). It provides many of the benefits of both independent and group projects, as well as a unique benefit of deeper involvement in other students’ projects. Although I have used it in Introductory Methodology and Psychometrics courses, it could be adapted for use in other methodology-based courses.

References


Notes

1. I thank Patti Connor-Greene for her helpful suggestions on an earlier draft of this article.

2. Send correspondence and requests for specific psychometric course assignments to Cynthia Pury, Department of Psychology, Clemson University, 418 Brackett Hall, Clemson, SC 29634–1355; email: cpury@clemson.edu.

Research Methods Textbooks: An Objective Analysis

Sherri L. Jackson
Susan M. Lugo
Jacksonville University
Richard A. Griggs
University of Florida

Research methods is one of the more prevalent and frequently required courses in the undergraduate psychology curriculum. To aid teachers in the textbook selection process for this course, we compared 26 methods textbooks published between 1995 and 1999. We analyzed the texts’ demographic qualities, use of pedagogical aids and illustrative material, and topic coverage. The variability observed in these analyses should help teachers to find the text that best suits their preferences and needs.

The curriculum committee at the St. Mary’s Conference on Enhancing the Quality of Undergraduate Education in Psychology (Brewer et al., 1993) reported that research methods courses are especially important to undergraduate psychology education because they “foster analytical skills,” “encourage critical thinking,” and provide skills that can “enhance lifelong learning” (p. 173). Messer, Griggs, and Jackson (1999) found that 73% of the 518 degree options in their undergraduate curriculum survey required a research methods course. Similarly, Perlman and McCann (1999a), in their frequency count study of undergraduate psychology course listings, found that not only was research methods one of the most frequently listed courses but also that it had increased more than 10% since 1975. In their companion study on the structure of the undergraduate psychology curriculum, Perlman and McCann (1999b) found that other than the introductory, statistics, and capstone courses, research methods was the most frequently required course for psychology majors. Additionally, in his “psychological literacy” study of important terms for psychology education, Boneau (1990) found that 42 of the “top 100” terms were from research methods–statistics. Clearly, research methods courses and their content are essential to undergraduate education in psychology.

Given their importance, articles concerned with research methods have been a mainstay in Teaching of Psychology. Johnson (1995) provided an index for articles published in Teaching of Psychology between the years of 1974 and 1994. More articles were published on research methods (83) than any other topic. In a recent content analysis of Teaching of Psychology from 1991 through 1998, Griggs, Jackson, Christopher, and Marek (1999) found that articles on research methods averaged about 1 per issue. Surprisingly, however, not 1 article from 1974 through 1999 provided comparative data on research methods textbooks. The primary goal of this study was to fill that void. Using analytic techniques similar to those employed in recent studies of textbooks for other major psychology courses (e.g., Christopher, Griggs, &
A Virtual Lab in Research Methods

Barbara A. Sommer
Robert Sommer
University of California, Davis

A hands-on lab for a lower division research methods course used an online format with Web page, Web forms, an e-mail listproc, and chat room. The virtual section received a higher rating for overall value than did the in-person labs. Students liked its convenience and flexibility. There were no significant differences in examination performance between students who took the online lab and 12 others who requested it, but could not be accommodated. Compared with the traditional course, more time was required in constructing assignments and communicating with students. A major advantage was eliciting responses from all students in contrast with the usual lab section.

With increasing student numbers and a geographic as well as age-diversified student body, the Internet provides an opportunity for exploring various combinations of traditional and online learning. Dietz-Uhler and Bishop-Clark (2001) described the benefits of both synchronous and asynchronous computer-mediated communication on subsequent face-to-face discussion. Kazmerski and Blasko (1999) showed the advantages of using a computer-based interactive program for learning observational research techniques. Maki, Maki, Patterson, and Whittaker (2000) found that students taking an online version of introductory psychology showed more content knowledge (based on exam performance) than did those in the traditional lecture class, although the latter expressed greater satisfaction with the course.

These positive findings suggest that Web-based learning may be a way of accommodating increasing student enrollments. The development of online materials is likely to involve considerable preparation time and resources. However, offering lab sections for large classes also requires extensive resources. Providing 1-hr per week 20-student lab sections in a class of 300 students means that faculty or teaching assistants are required for 15 hr a week in addition to the lecture time. If online instruction can reduce the number of personnel required, then its development might be worth the start-up cost. There is also a potential savings in classroom space.

This article describes the development and outcome of an online laboratory section for a lower division research methods course. The four-unit course enrolls 250 to 300 students every quarter with 3 hr of lecture and a required 1-hr laboratory per week with section size limited to about 15 students. The goal of the laboratory section is to provide students with hands-on experience in using the methods described in the lectures.

Materials and Procedure

The first author set up and conducted one of the sections as a virtual lab in Spring 2001 to compare with the 15 in-person sections. A Web site delivered content and assignments. The instructor created interactive forms using a Web authoring program (Adobe GoLive). Other suitable software packages include Macromedia, Dreamweaver, and Microsoft FrontPage. Course management software such as WebCT, Blackboard, and eCollege could also be used. The materials for the online lab are available at http://psychology.ucdavis.edu/sommerb/vlab/.

The underlying rationale was one of active learning and a multimethod approach. Students did both individual and group work, with an emphasis on the latter. The laboratory assignments covered the following topics: ethics, observation, library search, experimentation, interviewing, questionnaire construction, descriptive statistics, inferential statistics, and writing a report in the style prescribed by the American Psychological Association. The instructor provided clear and highly structured lab assignments. In the traditional format, these exercises occurred in sections that met 1 hr a week under the supervision of a graduate teaching assistant (TA) with some individual and group work done outside of class.

The TAs met weekly with the instructor and received instructions for implementing the assignments. The online section assignments were as much like the in-class ones as possible within the constraints of the Internet environment. At the beginning of the course, students had the option of signing up for a virtual rather than an actual laboratory section. As we were unable to accommodate all of the students, we randomly selected 15 from the 27 who requested the virtual option.

To facilitate collaborative work, students in the virtual lab used a chat room provided by the local campus system. The chat room was for student use only; the instructor did not participate. The aim was for the chat room to provide an analog of the outside-class meetings of students in the other sections. (If not available on a campus system, chat room space can be found at commercial sites such as Yahoo!; see http://chat.yahoo.com/)

The section instructor set up a listproc—an e-mail list for group communication. For example, after collecting the personal space experiment data, the section instructor posted the results and asked that each student comment to the e-mail list. The e-mail list also provided an opportunity for further online discussion.
All interactions between students and section instructor took place online. Table 1 lists the manner in which students completed assignments and communicated with the instructor. Some of the assignments required collaboration (see Table 1).

The e-mail burden was manageable. Messages ranged from a high of 53 the first week (mostly concerning admission into the online section) to lows of 8 and 7 in Weeks 6 and 8, respectively, when students turned in assignments via electronic form rather than e-mail. With regard to individual e-mail activity, the mean number of e-mail messages from the 15 students for the quarter was 16.5 with a median of 14. Two students produced the maximum of 32, and the low end was 6 messages.

As the students attended lecture 3 hr a week, they had the opportunity for face-to-face contact with each other. The instructor's reading of the e-mail and chat room records suggested that much of their contact was made online.

**Evaluation and Discussion**

**Student Perspective**

At the end of the quarter students in the lab sections connected with the course filled out anonymous evaluation forms. The results provided an overall rating of the (a) section leader, (b) value of the exercises, and (c) value of the section.

Table 2 shows the mean ratings of the section leader, value of the exercises, and value of the section. There were significant differences across the mean ratings of the section leaders—the virtual section instructor plus five TAs (multiple sections combined for each TA), $F(5, 185) = 5.81, p < .0001$.

As the rating of the section leader might have a halo effect on the ratings of the value of the exercises and the sections, we included leader rating as a covariate in subsequent analyses. The ratings of the section leader were correlated with the both the exercise ratings, $r(191) = .37, p < .0001$, and section ratings, $r(191) = .46, p < .0001$.

<table>
<thead>
<tr>
<th>Online Form</th>
<th>Listproc</th>
<th>E-Mail to Instructor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory Ethics 1</td>
<td>Ethics 2</td>
<td>Attitude scaling 2</td>
</tr>
<tr>
<td>Observation 1</td>
<td></td>
<td>Descriptive statistics 2</td>
</tr>
<tr>
<td>Observation 2 (group assignment)</td>
<td></td>
<td>(pair assignment)</td>
</tr>
<tr>
<td>Library</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experiment 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interviewing 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interviewing 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attitude scaling 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Descriptive statistics 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inferential statistics</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Listproc = E-mail discussion.

The differences in ratings of the exercises across sections were not statistically significant, $F(5, 184) = 1.12, p = .35$. The differences in ratings of the value of the section did vary significantly, $F(5, 184) = 5.54, p < .0001$. Taking into account the ratings of the section leaders, the mean for the value of the virtual lab section was significantly higher than each of the other sections ($p < .05$, least significant difference test).

Students in the virtual lab responded to open-ended questions about the section. Asked what they liked best, all commented on its convenience and flexibility with regard to time. Two students mentioned the advantage of more equal participation in discussion, "great for shy people like me." The appreciation of flexibility and convenience echoes findings from other studies of online courses (Maki et al., 2000; Pear & Novak, 1996).

On the negative side, two students commented on the "coldness" or impersonality of the Internet interaction: "Since I could never see or speak to the instructor, there was a very impersonal, and almost unfriendly environment which made it difficult to speak up." The theme of impersonality has come up in informal discussion with students on our campus. Many of the students have traditional views of the university and believe that they deserve the opportunity to attend formal classroom lectures given by professors. The belief persists even among those who acknowledge that they cut class on occasion, sit in the back row of the lecture hall, and rarely ask questions. Some view alternative class arrangements as less worthy of the elite institution.

With regard to overall performance in the course, the students in the virtual class did not differ from the rest of the class, nor did their scores differ significantly from those of 12 students who had requested the section but could not be accommodated. The fact that the students chose to take the online section could be a factor in the favorable evaluation. When comparing traditional and online versions of an introductory psychology course, Waschull (2001) received higher ratings from students who chose the online section versus those in the traditional one. Compared with students in the traditional section, those assigned to an online section gave her a slightly lower rating on each of the nine items on the evaluation form. We were not able to test the effect of random assignment versus choice.

**Instructor’s Evaluation**

In contrast with the general classroom situation, in which a few students can dominate the discussion and others do not participate, the online response either in the form of an e-mail list or bulletin board elicits greater participation. The second author has been using a personal space demonstration in classes for several decades (Sommer, 1999). The electronic feedback provided in the e-mail exchange following the personal space experiment was the most detailed and perceptive that he has ever received. The feedback becomes an archival record that has potential for providing insight into student learning.

Unlike the usual face-to-face interaction of groups working outside of class, students’ use of the chat room resulted in a transcript of their discussions. In one instance, a group had
come up with the wrong solution to a problem in data analysis. By examining the transcript of their discussion, the instructor was able to see where their reasoning went awry.

On the negative side, there was a considerable amount of preparation time for the instructor—more than would be the case if one simply had to show up for an hour-long section meeting. This was the first time we offered the virtual lab and subsequent preparation time would be considerably less, depending on the degree to which we modified assignments.

There is also the time demand of e-mail. We do not know how often students contacted their section leaders by e-mail, but there certainly was more contact from the virtual lab students than had been the case when the instructor was lecturing to the class of 300. Managing queries from 15 students was not onerous, but it would be challenging with 150.

An e-mail filter can reduce the time load by directing the class-related e-mail messages to a single folder or directory where the instructor can handle them at a convenient time. Using a database for downloading student responses is another way of saving time. Our institution has a database program that saves the results from Web forms (e.g., students' responses) in tab-delimited files. These files are downloaded and read into a database or spreadsheet program such as Microsoft Excel. (See http://webtools.ucdavis.edu/dbtool/ for information. The program can be made available to interested institutions.)

As did some students, some instructors will find online communication to be cold and impersonal. On the other hand, here are two quotations from the e-mail listproc discussion of the ethics assignment:

It's such an eye opener to hear all the other responses. People brought up some excellent points that I would have never thought of myself. I guess that's why there are many people on IRB's [sic] and not just one. I also think that maybe if I would have heard other people's opinions and concerns before I sent my ideas, then my answers may have been different. Hearing other people's point of view puts a new perspective on my own.

To put my two cents in here, I agree that it was really interesting to get the opportunity to read everyone else's case responses. I think this virtual section will provide an opportunity to hear more differing points of view. In a physical discussion section, there is so often 3–4 people with strong (and loud) opinions that end up dominating the hour while the quieter people may never share. I appreciate the democracy of this format.

An instructor can take steps to reduce the impersonality, for example, by scheduling a face-to-face meeting at the beginning of the course, putting photographs of the instructor and students online, and personalizing Web materials and e-mail messages. A discussion about the perceived impersonality of online communication may head off potential alienation. Presenting the lab online was successful and can be replicated with easily made modifications to suit a particular curriculum.

References


Note

Send correspondence to Barbara A. Sommer, Psychology Department, University of California, Davis, One Shields Avenue, Davis, CA 95616–8686; e-mail: basommer@ucdavis.edu.

Table 2. Mean Ratings and Standard Deviations of Value of the Section Leader and Adjusted Mean Ratings of the Overall Value of the Exercises and of the Section

<table>
<thead>
<tr>
<th>Section(s)</th>
<th>Rating of Section Leader&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Value of Exercises&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Value of Section&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>Virtual lab</td>
<td>1.71</td>
<td>.91</td>
<td>1.79</td>
</tr>
<tr>
<td>TA 1</td>
<td>1.88</td>
<td>.61</td>
<td>2.33</td>
</tr>
<tr>
<td>TA 2</td>
<td>1.34</td>
<td>.57</td>
<td>2.10</td>
</tr>
<tr>
<td>TA 3</td>
<td>2.01</td>
<td>.74</td>
<td>2.29</td>
</tr>
<tr>
<td>TA 4</td>
<td>1.53</td>
<td>.64</td>
<td>2.09</td>
</tr>
<tr>
<td>TA 5</td>
<td>1.37</td>
<td>.63</td>
<td>2.11</td>
</tr>
</tbody>
</table>

Note. TA = teaching assistant. <sup>a</sup>Scores ranging from 1 (excellent) to 5 (terrible). <sup>b</sup>Scores ranging from 1 (very valuable) to 4 (not valuable).
Applying Technology to Facilitate Poster Presentations

Pam Marek
St. Michael’s College

Andrew N. Christopher
Anderson College

Cynthia S. Koenig
University of Florida

Because college graduates with baccalaureate degrees in psychology are often employed in positions that are not directly related to psychology, we offer a rationale for introducing applied technological skills in the classroom. We focus specifically on presentation software, primarily as a vehicle for creating posters. Given the surge in popularity of poster presentations and the impact of high-quality visual displays, we review sources of information for preparing posters and graphs. We also offer guidelines for the use of PowerPoint™ presentation software to create single-unit poster presentations, a skill applicable in a variety of academic and employment settings.

Students clearly shoulder part of the responsibility for their lack of preparedness. Yet, it seems reasonable to propose that the responsibility of psychology faculty, as educators, should extend beyond the dissemination of psychological knowledge. To varying degrees, psychology instructors are already incorporating the development of important employment-related competencies into classroom experiences and requirements. Traditionally, the umbrella of applied skills has encompassed activities to enhance competencies linked to oral and written communication, analytical ability, and interpersonal relationships. Recently, computer expertise has been added to the list. For example, technological ability has been listed as an objective for psychology majors (Allen, Noel, Deegan, Halpern, & Crawford, 2000) and among those skills that employers find desirable in psychology graduates (Landrum, Davis, & Landrum, 2000). Additionally, Lloyd (1997a, 1997b) indicated that research skill, including the ability to communicate results via tables and graphs, is a valuable asset for graduates preparing for entry-level positions in psychology (e.g., behavior analyst), business (e.g., sales representative), and other fields (e.g., college admissions recruiter).

We see the psychology classroom as an appropriate arena for providing educational experiences that promote technological, research, and communication skills relevant to an array of potential careers. Modeling the use of technology for presenting substantive content and for training students in the application of technology exemplify two such educational experiences. In this article, we discuss the application of one specific technology—presentation software. After briefly commenting on its use in the classroom, we narrow our focus to poster-based communication of research results and the use of presentation software to facilitate poster preparation, a competence applicable in a variety of settings.

Use of Presentation Software

Cognitive principles suggest that a coherent blend of verbal–visual material boosts retention of scientific concepts beyond the level attained by verbal summary alone. Mayer, Bove, Bryman, Mars, and Tapango (1996) demonstrated such a benefit on retention of textbook material. The increasing availability of computers in classrooms may encourage instructors to aim for similar benefits from classroom lectures by using presentation software to readily integrate verbal–visual displays into lecture-based courses. Seaman (1998) provided guidelines for doing so, illustrating how faculty could use presentation software to construct displays to serve an array of instructional purposes (e.g., advance organizers, graphic representation, review, inquiry).

As students increasingly search for relevance in their classroom experiences, they may quickly realize the potential
value of presentation software for their personal use. For example, students in the first author’s research methods courses, motivated by an interest to master a new medium modeled by the instructor, have been enthusiastic about using presentation software. Students have voluntarily employed presentation software to present proposals, expose experimental participants to a variety of stimuli (e.g., faces with different expression), and communicate their research findings. Similarly, students in the first author’s social psychology course have chosen presentation software over more traditional transparencies as a vehicle for communicating research-oriented presentations to their classmates, also noting how it eases the task of preparing handouts. Thus, echoing positive student ratings of technology-assisted lectures (Adams, 1998), anecdotal evidence has suggested that modeling of and instruction in the basics of PowerPoint™ enhances student interest in the development and organization of oral reports. Moreover, given that increased experience with technology appears to be a key factor in reducing technophobia (Scott & Rockwell, 1997), guiding students through the basics of presentation software may provide an impetus for them to apply it in future classroom and career-related tasks.

Poster-Based Communication of Research Results

Students seem particularly intrigued by the potential of this technology for creating well-organized and colorful posters. As a means of communication, posters have surged in popularity (Woolsey, 1989), emerging not only at conventions and teaching exchanges (Matthews & Jacobs, 1986), but in classrooms as well (Baird, 1991; Mansfield, 1993). In the classroom, poster sessions encourage students to interact during the preparation stage, providing an opportunity for collaborative learning, and during presentation, providing a forum for inquiry and discussion (Baird, 1991). In a convention setting, attendees’ opinions of posters are influenced as much by visual presentation as by content (Welch & Waehler, 1996). Specific comments about visual presentation often pertain to high-quality graphics and use of color as well as large print within the title and body of the presentation. To enhance the effectiveness of poster-based communication, Woolsey emphasized that poster design should incorporate visual differentiation of key and subsidiary ideas and suggested that color can either unify or differentiate various sections of data (see also Seaman, 1998).

To assist students and instructors in designing posters to communicate research findings, general guidelines (e.g., Bennett, 2000; Martin, 2000; Szuchman, 1999) have emphasized how to identify and select the most important material for posters, avoiding the inclination to include the bulk of a written report. The guidelines also have cautioned presenters to prepare textual material to ensure that it is readable at a distance of 3 ft. The actual suggested font size depends on whether the poster includes several 8½ × 11 sheets, individually printed on a standard printer, or is a single-sheet poster, enlarged for printing on a specialized wide-format printer. The guidelines also include suggestions regarding column layout, use of bullets to minimize the size of blocks of text, and preparation of handouts for distribution at the presentation site. All guidelines have accentuated the importance of using graphical material but have not covered the basics of graph preparation.

To fill this gap in coverage, Pittenger (1995) highlighted specific information about maximizing the accuracy and effectiveness of graphs, applicable to both report and poster presentation. Pittenger clearly distinguished between nonsential graph elements, or chartjunk, and the appropriate design of an effective graph, noting, for example, that two-dimensional graphs are more readily interpretable than are three-dimensional graphs. He also provided a list of selected texts and chapters on graphing techniques, commenting on the contents of and examples in each. Because graphs often command a prominent position on posters, this article is a valuable addition to the more general poster preparation guidelines.

Using Presentation Software to Facilitate Poster Preparation

To supplement existing information about poster design and preparation of graphs, we have prepared specific step-by-step guidelines for creating a poster on a single slide using PowerPoint. We initially prepared guidelines to fill requests from audiences who viewed our single slide PowerPoint posters at recent conventions. Our experience suggests that, compared to traditional modes of poster preparation, PowerPoint facilitates alignment, arrangement, and coloring of poster components and insertion of relevant illustrative material (e.g., graphs, tables, and photos). Concurrently, it eliminates the need for a "toolbox" of cutting and pasting supplies and reduces assembly time at the convention or presentation site. Finally, because printed copies of letter or legal size versions of the poster are readily available, use of PowerPoint may substantially simplify preparation of handouts for interested viewers. Printing the poster requires a wide format full-color printer, often available at nationwide printing facilities. At first glance, the cost at such noninstitutional providers (approximately $9 per square foot) may appear prohibitive, particularly in relation to a conventional poster. However, considering the cost of poster board, backing for individual items, adhesive, and professional services for cutting poster board, the incremental cost of a PowerPoint poster appears justifiable, particularly given its potentially greater impact on audiences.

Concluding Comments

The authors are newcomers to the use of this technology; yet, we found the software to be generally user friendly. As with any new venture, it took time and patience to initially master the intricacies of formatting and arranging poster content to enhance visual appeal. However, the subsequent positive reactions of viewing audiences provided a substantial reward. Moreover, the ease of setting up a
one-piece poster display at convention sites has been a welcome convenience. Given these advantages, we invite interested readers to review or download a copy of our guidelines for preparing a poster on a single PowerPoint slide (Marek, Christopher, & Koenig, 2001) from the Office of Teaching Resources in Psychology Web site (http://www.lemoyne.edu/OTRP/index.html).

Our students have successfully used these guidelines to create posters that they presented at regional conferences. Moreover, we suspect that our students’ increased familiarity with this technology will benefit them both in graduate school and work environments. From a broader perspective, applying technology to facilitate poster presentation is but one way of conjoining topical coverage with applied skills to scaffold student growth within and beyond academia. We consider that such integration adds an incremental benefit to the conceptual knowledge we aim to convey. In short, we believe that incorporating technology into psychology courses has benefits beyond technology’s potential to increase student interest in topical material.

References


Notes

1. Pam Marek is now at Anderson College. Andrew N. Christopher is now at Albion College.

2. Portions of this article were presented as a poster at the Southeastern Conference on the Teaching of Psychology, February 2000, in Kennesaw, GA.

3. We thank Joann Benigno, David Pittenger, and three anonymous reviewers for their comments on earlier drafts of this manuscript.

4. Send correspondence to Pam Marek, Department of Psychology, Anderson College, 316 Boulevard, Box 1166, Anderson, SC 29621–4035; e-mail: pammarek@yahoo.com.
Converting an Experimental Laboratory Course from Paper and Pencil to Computer
Maryellen Hamilton and Lisa Geraci
Teaching of Psychology 2004 31: 141
DOI: 10.1207/s15328023top3102_7

The online version of this article can be found at:
http://top.sagepub.com/content/31/2/141

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Society for the Teaching of Psychology

Additional services and information for Teaching of Psychology can be found at:

Email Alerts: http://top.sagepub.com/cgi/alerts
Subscriptions: http://top.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Apr 1, 2004
What is This?
Converting an Experimental Laboratory Course From Paper and Pencil to Computer

Maryellen Hamilton
Saint Peter’s College

This article provides suggestions for developing a laboratory-based research methods course using computers. We describe important considerations for creating this type of course including selecting software, choosing experiments, and teaching students with different levels of computer skill. We also include 3 model projects that required increasing methodological and programming knowledge. The ultimate goal of the course was to have students submit a final project that they designed, ran, and programmed independently.

It has been our experience that most undergraduate students find their experimental laboratory course to be among the most demanding courses they take as part of their psychology major. Traditionally, professors teach this course in a lecture classroom, with students conducting experiments using paper and pencil. However, given that many experiments (particularly those on human cognition) are conducted on computer, it is important to train students to use computers to conduct research. Others have discussed the possibility of enhancing research methodology courses through the use of computers (e.g., Goolkasian, 1985; Peden, 1987; Rittle, 1990); however, this article provides a specific template for creating such a course including experiment and software suggestions. We derived this template from our experience developing and teaching a computer-based research methods course in human cognition at the State University of New York at Stony Brook. Although the class we developed was a research methods course in human cognition, most of the suggestions are applicable to a general laboratory course on research methods.

We describe three main considerations for developing a computer-based experimental laboratory course: selecting software, choosing experiments, and teaching students with different levels of computer skill. We provide solutions that attempt to enrich students’ learning of research methods. Our goal for the course was to have students design and program their own independent project on the computer. We required students to program and conduct a series of three experiments prior to completing an independent project. Each experiment increased in theoretical and methodological complexity as well as in programming difficulty. For each successive experiment, students wrote additional sections of an American Psychological Association (APA)-style paper. By the third project they wrote an entire paper.

General Course Considerations

There were three main considerations in developing our computer lab course. First, we had to make software and hardware decisions. We needed experimental software, a program for data analysis, and a way to make stimuli available to each student. We chose Cedrus’s Superlab (Abboud & Sugar, 1997) for our experimental software because the program is easy to learn. Superlab is a Windows®-based program that allows experimenters to present stimuli (pictures or words) on a computer screen and record participant responses. The program records accuracy and reaction time for each response and stores the data in a file that students can open using a variety of spreadsheet programs or statistical packages. For example, using a spreadsheet such as Microsoft Excel®, students can sort data, obtain means, and perform data analyses. Students used a private network drive accessible with a password to make changes to their experiment and to continue working on it in subsequent classes. Their network space also gave them a place to save their work for grading. In addition, the entire class could access a network drive where we placed all the materials for each experiment. Although the students had access to this drive, they could not edit it.

Second, we had to select experiments that would work in a classroom setting. Conducting experiments in the classroom required that our students act as both the experimenter and participant in a single study. This decision meant that we had to select experiments with effects that would be obtained even when participants knew of the purpose of the experiment. Also, we had fewer than 20 students in each lab section. Therefore, we had to select experiments with effects large enough to be obtained with a small sample size.

Third, we had to consider that students varied considerably in experience using computers. Some of our students had only word processing experience, whereas others were computer science majors. Therefore, we structured the course around this diversity in computer experience. To aid in this endeavor, the second author wrote a manual specifically for the course.1 The manual provided novice users with

1We encourage professors who want to teach this class to design a manual that is specific to their course needs (e.g., student’s abilities, software specifications, course emphasis).
detailed instruction, without forfeiting class time for the more experienced users. Using this manual, students with little computer knowledge were typically able to program simple experiments within a few classes. Another way we dealt with the variability in computer experience was to have students work in pairs. Although students had a computer to themselves, partners could help each other solve programming issues. We rotated partners across each of the three projects so that everyone had an opportunity to work with a more sophisticated computer user.

Specific Course Projects

We selected projects with programming requirements that allowed us to explore specific methodological issues. These selections were important because our goal was to teach research methods as well as computer skills. We used programming specifications to teach students about experimental precision and to make explicit critical research design decisions. For example, certain critical points in programming required students to think about whether the design they were using was within or between subjects. Students needed to make choices about how many lists to program for counterbalancing. The program also required them to code the stimulus presentations (e.g., as studied vs. nonstudied or picture vs. word), which forced them to declare their independent variables and learn about the levels of a variable. We highlighted (both in lecture and in the manual) these issues and presented various programming options to encourage students to think about these design decisions as they programmed. We detail three projects that increase in methodological, theoretical, and programming complexity and required increasingly sophisticated data analyses.

For their first experiment, we chose the classic Stroop (1935) experiment because it met the criteria of having a large effect that was relatively immune to participants’ (our students) knowledge of the hypothesis being tested. In addition, this experiment has a simple design with one independent variable with three levels (Stroop, match, and neutral) and one dependent variable (reaction time). Using this experiment as a model, we introduced the following research concepts: hypothesis testing, independent and dependent variables, within- versus between-subject designs, counterbalancing, and the use of ANOVA with post hoc analyses. This experiment required minimal programming because we preprogrammed all the conditions except for the simplest condition, the match condition (where the word matches the color in which it is printed). Each student received a copy of our manual that included a step-by-step procedure for how to program the match condition. The computer skills introduced for this first experiment were creating text stimuli, timing stimuli, randomizing trials, and manipulating and analyzing data using Excel. Students submitted an APA-style report including a title page, abstract, method section, table, and printout of the data analysis.

For the second project, students attempted to replicate a divided attention experiment with ambiguous figures (Reisberg, 1983). The experiment examines why people first see one interpretation and then the other when looking at an ambiguous figure. Whereas in the previous experiment there was only one predicted hypothesis, this experiment had three competing hypotheses regarding the role of attention in mediating this effect. Students tested these hypotheses by comparing the amount of time it took for ambiguous figures to change from the first to the second interpretation under both divided and full attention conditions. This experiment met the criteria previously outlined (i.e., a large effect that could occur with informed participants) and introduced the following: competing hypotheses, confounds, outliers, and the use of t-tests. It also required minimal programming, which was important because it was the first experiment that students programmed on their own. The design of the experiment was simple and few stimuli were needed. We placed two sets of ambiguous figures (provided in Reisberg, 1983) on a network drive. Half of the class used one set and the other half used the other set so that they were not tested on the same stimuli that they programmed. Students presented one stimulus under divided attention and the other under full attention and created different versions of the experiment to counterbalance the items and the order of conditions. We introduced the following computer skills: creating a new experimental file, accessing stimuli on the network, ending trials with a key press to record reaction time, and creating an instruction screen. For this project each student submitted an APA-style report that included a title page, abstract, method, and results section with data analysis and figure.

For the final project, students conducted a conceptual replication of Weldon and Roediger’s (1987) experiment, which demonstrated that the picture superiority effect (better memory for pictures vs. words) can be reversed depending on the cues at retrieval. This experiment had a more complicated design than the previous experiments, two independent variables, and predicted a crossover interaction. Participants studied pictures and words and took either an explicit word fragment cued recall test or a recognition test. Students programmed this entire experiment themselves. So that no one would be tested on the same material that they had spent time programming, we provided students with two different sets of materials on the network drive. We introduced the following research concepts in the third experiment: 2 × 2 mixed factorial design, counterbalancing for study status and encoding format, main effects and interactions, informed consent, and debriefing. Computer skills introduced in this project included creating study and test files; creating a fixed order for events; coding trials (picture vs. word, studied vs. nonstudied) and coding responses as either correct or incorrect; and sorting data by hits, misses, false alarms, and correct rejections. Unlike the previous two projects, we provided the students with a copy of the Weldon and Roediger’s paper to use as a model because this was the first time students wrote an introduction to their APA-style paper. Students submitted a full APA-style report including title page, abstract, introduction, method, results, and discussion, with data analysis and figure.

Conclusions

For the final project, each student independently designed and programmed an original experiment. They gathered their stimuli and created their consent and debriefing forms. Stu-
Students tested a minimum of 20 participants and performed their data analysis. The final result was a full APA-style report and a conference-style presentation with overheads. Forty-six of the 48 students (across three sections) completed the projects successfully. Student projects varied greatly in design and programming complexity. Several of the programs were extremely complicated (including audio), and many students won school research awards for their work in this class.

Students reported on the evaluations for the class that they had learned a great deal in the course and would recommend the course to others. On a scale from 1 (strong agreement) to 7 (strong disagreement), the students' mean overall satisfaction with the course was 1.81 (SD = 1.04), which was better than the university average of 2.08. When asked how they agreed with the statement, “They would highly recommend this course to others” the mean was 2.14 (SD = 1.01). In addition, many of the evaluations included comments that the students “learned more in this course than in any other course they had taken in college.”

By introducing programming to this course, we believe students were better prepared for graduate work in psychology because they could use computers to implement experiments and analyze their data. We also believe that, by adding the programming component, students thought critically about research design issues (i.e., the experiment will not run appropriately if the design is not precisely specified) and learned to conduct extremely careful experiments.

References


Notes

1. A preliminary version of this article was presented at the seventh annual APS Institute on the Teaching of Psychology, Miami Beach, FL, June 2000.
2. Send correspondence to Maryellen Hamilton, Department of Psychology, St. Peter’s College, 2641 Kennedy Boulevard, Jersey City, NJ 07306; e-mail: mhamilton@spc.edu.
Using CHAINS, a QuickBASIC 4.5 Program, to Teach Single-Subject Experimentation with Humans
Marshall Lev Dermer
Teaching of Psychology 2004 31: 285
DOI: 10.1207/s15328023top3104_9

The online version of this article can be found at:
http://top.sagepub.com/content/31/4/285

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Society for the Teaching of Psychology

Additional services and information for Teaching of Psychology can be found at:

Email Alerts: http://top.sagepub.com/cgi/alerts
Subscriptions: http://top.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Oct 1, 2004

What is This?
Students enrolled in a single-subject design course studied the repeated acquisition of response sequences by using CHAINS, a QuickBASIC 4.5 program, which runs in DOS or Windows®. For about 2 months, students examined the learning of such sequences as a function of various treatments. Each week students graphed their data, discussed their research, modified their experiments, and then wrote American Psychological Association-style manuscripts. Students positively evaluated CHAINS and the instructional process. Suggestions for future instruction include distributing students' manuscripts the following semester so new students can replicate earlier work and using the alternating treatments design, which quickly permits comparing treatments.

Although psychological practitioners most often help individuals, psychological researchers most often aggregate data across individuals. Such aggregation, however, may obscure how treatments affect an individual. Single-subject experimentation is important because it can reveal whether and how treatments affect an individual’s performance (Dermer & Hoch, 1999). Indeed, researchers in behavior analysis have conducted single-subject experiments that have contributed to a number of highly general technologies for working with nonhumans (e.g., Pryor, 1997) and humans (e.g., Lovaas, 1993).

Despite the utility of single-subject experimentation and recent growing interest in such research with humans (Dermer & Hoch, 1999; Morgan & Morgan, 2001), there is a dearth of instructional software. Silva (1999) introduced software for the Macintosh® operating system that allows students to study shaping. The software tracks how a student moves a cursor on a computer screen and shapes movements as a function of various parameters until the cursor “hits” a target (Silva, 1999; Silva, Yuille, & Peters, 2000).

For the DOS and Windows® operating systems, Dermer and Dermer (2000) introduced software that allows researchers to study the effects of various treatments on human learning and performance. The software, CHAINS, implements the repeated acquisition of chains procedure. The participant’s basic task is to depress a sequence of numeric keys (the chain). Depressing a correct key places the corresponding sequence during the acquisition components is constant across sessions.

For research purposes, Dermer and Dermer (2000) provided QuickBASIC 4.5 source code, instructions for modifying the software, and a compiled demonstration program. For instructional purposes, I compiled four versions of CHAINS. Each version presents 90-sec performance and acquisition components. Each component is signaled by a colored bar that appears on the screen’s left side. For all components the sequence length is constant (e.g., seven keys). For each component CHAINS calculates the proportion of errors (the total number of incorrect responses divided by the total number of responses) and rates per minute of correct and incorrect responding (see Dermer & Dermer, 2000).

By examining these measures as a function of experimental treatments, a researcher can determine whether the experimental treatments affected entering “new” sequences (acquisition data) and entering a “well-learned,” highly practiced sequence (performance data). Additionally, a researcher may compare data from these two components because performance data may reflect various uncontrolled variables such as the participant’s motivation and alertness whereas acquisition data reflect these uncontrolled variables plus learning.

Researchers may use a number of experimental designs. In the ABAB design, a researcher continues a treatment for many sessions until responding achieves a “steady state,” that is, from session to session responding assumes a characteristic level and tends to neither increase nor decrease across sessions (see Baron & Perone, 1998; Perone, 1991) once responding achieves a steady state, the researcher withdraws the treatment and later reintroduces it after responding has stabilized during a comparison treatment (Barlow & Hersen, 1984, chap. 5). The researcher assesses treatment effects by comparing responding during steady states. In contrast, a researcher using an alternating treatment design can circumvent such a lengthy, steady-state analysis. In the alternating treatment design, the researcher changes treatments from session to session or within sessions (Barlow & Hersen, 1984, chap. 7). The researcher assesses treatment effects by comparing levels of responding given a background of response instability. Next, I describe how my students used these two
designs to examine the effects of various treatments on responding during the performance and acquisition components of CHAINS.

Method

Participants and Setting

Nine students who had completed a laboratory course in research methods in psychology participated as part of their work for a course on single-subject design. The course met once weekly for 2.25 hr. I used Barlow and Hersen (1984) as the primary text. All students conducted and documented a single-subject experiment of their own design. Eight students chose to use CHAINS.

Materials

Software. Four versions of CHAINS (see Dermer & Dermer, 2000) presented four, 90-sec components: performance, acquisition, performance, acquisition. The performance component used a constant numeric sequence from session to session; the acquisition component used numeric sequences that varied across sessions but were constant within a session. Each version of CHAINS presented 7-, 8-, 9-, or 10-digit sequences (e.g., CHAIN147.EXE presents 7-digit sequences). The four versions are in ChainsForClass.EXE, available at ftp://ftp.csd.uwm.edu/pub/psychology/BehaviorAnalysis/software/educational/chains-dermer/. The downloaded, self-extracting file creates the directory C:\CHAINS and loads supporting files. Before activating a program, a 3.5-in. disk should be inserted into the A drive where CHAINS creates data files (cf. Dermer & Dermer, 2000).

Evaluation. Students evaluated instruction using a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The evaluative statements were: (a) “CHAINs was easy to install on a computer,” (b) “CHAINs was easy to use,” (c) “Using CHAINS enhanced my understanding of the process of conducting a single-subject experiment,” (d) “For many weeks, students presented and discussed data collected using CHAINS. These presentations and discussions enhanced my understanding of the process of single-subject research: designing, implementing, and evaluating a single-subject experiment,” and (e) “I recommend using CHAINS in future single-subject design classes.”

Procedure

Students first read and discussed an article that discussed various misconceptions about single-subject experimentation (Dermer & Hoch, 1999) and an article about using CHAINS (Dermer & Dermer, 2000). Next each student designed, conducted, and documented a single-subject experiment in which the student or another student participated. Students collected data at least 5 days per week. In their subsequent American Psychological Association-style reports, students emphasized operations rather than theory. Finally, students submitted the disk that contained the data files that CHAINS generated and copies of these files they may have printed.

For about 2 months, students conducted experiments, graphed data, and converted graphs to overhead transparencies for class presentations. Each presentation initially required about 12 min, but as the semester progressed 5 min usually sufficed. At the semester’s end, five students anonymously responded to the evaluation. Two more students responded to mailed evaluations after course grades were assigned.

Results and Discussion

Most students first used some version of the ABAB design but because of limited time their data did not satisfy conventional stability criteria (see, e.g., Baron & Perone, 1998, pp. 49–57). One student, using an ABA design, did collect rather clear data. Her experiment assessed the effects of the program’s auditory feedback, which signals many events including correct and incorrect responding. She aggregated data across the two presentations of the performance and acquisition components to calculate the overall proportion of erroneous responses for each component. Figure 1 shows these proportions as a function of the component in effect and the presence or absence of auditory feedback. For the first nine sessions, the student used the program’s auditory feedback. For the next nine sessions she discontinued auditory feedback. This change correlated with an increase in the proportion of erroneous responses for the performance and acquisition components. The increase was somewhat greater for the acquisition component than for the performance component. Additionally, discontinuing feedback correlated with an increase in between-session variability, particularly for the acquisition component. For the last four sessions the student reinstated auditory feedback. This change correlated with an immediate reduction in the proportion of erroneous responses. The reduction was greater for the acquisition component than for the performance component. Reinstating feedback also reduced between-session variability. The entire pattern of changes suggests that discontinuing auditory feedback had two effects: It increased the proportion of erroneous responses and increased between-session variability, particularly for the acquisition component.

In behavior analysis it is customary to further examine data to determine whether other levels of analysis reveal systematic relations. CHAINS calculates a variety of measures to facilitate such an examination. Accordingly, on a component-by-component basis, the student graphed the rates per min of correct responding and separately graphed the rates per min of incorrect responding. These two graphs clarified how these more fundamental measures contributed to the overall proportion of erroneous responses in Figure 1.

Most experiments, however, have not been so definitive. Over several semesters, the treatments that have not affected

1Janet Adornato collected these data during Spring 2001 and deserves special thanks.
responding include a fatty-acid food supplement that purportedly aids memory and playing or listening to relaxing guitar music or exercising before sessions. Whether a treatment affects responding depends on program parameters. A critical parameter is sequence length. Most students used short sequences but longer sequences may be more sensitive. Students should be encouraged to explore these parameters. Such tinkering enhances discerning cause and effect relations at the level of the single participant.

As the semester progressed most students switched from ABAB designs to alternating treatments designs. Review of Figure 1 indicates that an alternating treatments design is embedded within the ABA design. For every session, CHAINS twice presented the performance and acquisition components. As previously noted, during the performance component one sequence is used across sessions whereas during the acquisition component a sequence is unique to a session. Figure 1 indicates that the acquisition component nearly always produced more errors than did the performance component. Moreover, this effect was evident in 9 days despite responding varying from session to session. Although the effect may be limited to the components being presented in a fixed order, the result indicates the utility of the alternating treatments design for quickly comparing treatments given response instability from session to session.

On the evaluation, students reported CHAINS as easy to use and the related instructional activities valuable. All but one student strongly agreed with the four statements assessing ease installation, easy use, enhanced understanding of the process of single-subject experimentation, and recommending future use of CHAINS. All students strongly agreed with the statements that the weekly presentations of data and discussions were useful. Some students also provided comments, for example:

Chains was an invaluable learning tool. I cannot imagine how a textbook or journal article could have been a substitute for the experience of running sessions, graphing data, and deciding what treatment to implement. Also the program, being more or less foolproof, only enhanced the value of CHAINS as an educational tool.

These positive evaluations may have been due to the small class size but during the semester students often mentioned enjoying the freedom of selecting treatments to investigate as well as designing and modifying their research as they collected data. This “tinkering” with the experiment or “investigative play” (Hayes, Barlow, & Nelson-Gray, 1999, pp. 134–136) was absent from the highly structured exercises students completed for their course in research methods in psychology.

There are several ways to further improve instruction. Providing students with reports of experiments conducted during previous semesters would help them document their experiments and encourage them to replicate or refine previous work. Replication and refinement are vital activities that deserve classroom support (for a discussion of the critical role of replication in Pavlov’s laboratory, see Babkin, 1949, pp. 70–72, 112; Campbell, 1969, p. 367). Instruction may be further improved by encouraging students to use alternating treatments designs when comparing fast-acting treatments of short duration. Also, although students in my course usually scheduled only one or two sessions per day, multiple sessions may be completed in a short time if students schedule breaks between sessions.

Although CHAINS is a simple program, its use as an instructional tool may evoke apprehension among instructors new to single-subject research. These instructors may, therefore, want to review introductory texts on single-subject research (e.g., Bailey & Burch, 2002; Barlow & Hersen, 1984) and advanced discussions (Baron & Perone, 1998; Perone, 1991) and perhaps review the literature on repeated acquisition procedures (e.g., Cohn, MacPhail, & Paule, 1996; Cohn & Paule, 1995).

Instructors, however, can best allay their apprehension by conducting single-subject research with CHAINS. For example, I conducted an 11-day experiment in which I collected data at different times during my sleep–wakefulness cycle (about 3.5 hr after arising vs. just before retiring; Dermer & Dermer, 2000). Alternatively, instructors could easily replicate the experiment that manipulated auditory feedback, particularly if they used an alternating treatment design that either alternated or randomized the presence versus absence of auditory feedback for blocks of two sessions. By using CHAINS and analyzing their data, instructors would gain expertise and confidence that they could share with their students. Also, for their research methods courses, these instructors would best know how to design and introduce their new “hands-on” unit on single-subject research.

References


Notes

1. I thank the students in my single-subject design courses, during the Spring of 2001 and 2002, for using CHAINS and providing constructive suggestions. I also thank Noah Dermer, Robert Hessling, and Susan Lima for reviewing earlier versions of this article.

2. Send correspondence to Marshall Lev Dermer, Department of Psychology, University of Wisconsin–Milwaukee, Milwaukee, WI 53201; e-mail: dermer@uwm.edu.
Virtual Instruction: A Qualitative Research Laboratory Course
Lee M. Stadtlander and Martha J. Giles
Teaching of Psychology 2010 37: 281
DOI: 10.1080/00986283.2010.510971

The online version of this article can be found at:
http://top.sagepub.com/content/37/4/281

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Society for the Teaching of Psychology

Additional services and information for Teaching of Psychology can be found at:
Email Alerts: http://top.sagepub.com/cgi/alerts
Subscriptions: http://top.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Oct 1, 2010
What is This?
Virtual Instruction: A Qualitative Research Laboratory Course

Lee M. Stadtlander and Martha J. Giles
Walden University

Online graduate programs in psychology are becoming common; however, a concern has been whether instructors in the programs provide adequate research mentoring. One issue surrounding research mentoring is the absence of research laboratories in the virtual university. Students attending online universities often do research without peer or lab support and without previous experience in supervised research projects. This article describes a virtual qualitative research lab course at an online university that both mentored students and collected data through a virtual qualitative lab on a national scale. The authors also address logistics, problems and issues that arose, and suggestions for future courses.

There are currently a number of graduate psychology programs that are predominately online (e.g., Capella University, Fielding University, and Walden University). However, a key element of graduate training is the in-depth intellectual mentoring, particularly in research training. There has been concern as to the ability of instructors to mentor students in an online environment (Belar, 2006). This article explores the implementation of a graduate online qualitative laboratory and its effectiveness in terms of research methodology and in mentoring the student participants.

Before the 1960s, a research-focused model leading to the PhD was typical (Ellis, 1992; Roe, Gustad, & Moore, 1959). During this time, students were trained to teach and do research with the expectation that they would enter into academic positions (Altman, 1990; Ellis, 1992).

In the 1960s, as more career options became available for clinical psychologists, the traditional model began to be questioned (Ellis, 1992). However, the issue of research mentoring remains contentious, particularly for future practitioners (Gelso, 2006). Opportunities for research go beyond simply learning research skills, but training also needs to influence students at the motivational level. As Gelso (2006) described, “we need to light a fire under our students—to show them how research can be exciting and rewarding” (p. 4).

Research and Graduate Education

The importance of research mentoring in graduate education has long been a concern for psychology.
(i.e., blackboard, e-campus), as well as telephone communication and in-person meetings.

Some issues are unique to the online environment. As Stadtlander (1998) highlighted, personality issues can become an issue due to the lack of social context cues (Hiltz & Turoff, 1978; Rice & Love, 1987). Dixon and Stone (1997 as cited in Rudestam, 2004) reported that affective messages strongly influence online students and such messages tend to be read and reread. Therefore, it becomes important for participants to carefully write postings so that they are not misinterpreted (Rudestam, 2004).

Online Research Labs

There have been a few papers on virtual research laboratories; however, typically the use of shared remote equipment was the basis of the paper. For example, González-Castaño et al. (2001) designed an Internet access laboratory that provided remote access to equipment used in a computer architecture laboratory. Hesselink et al. (2003) introduced a design for a remote laboratory on an optical processor. Duan, Hosseini, Ling, and Gay (2006) detailed a lab-based computer systems course. However, there are no descriptions of psychology-based virtual research labs in the literature.

Qualitative Research

Qualitative research methodology refers to research that produces descriptive data, people's own written or spoken words, and observable behavior (Taylor & Bogdan, 1998). Qualitative researchers are concerned with the meanings people attach to issues in their lives. Researchers develop concepts, insights, and understandings from patterns in the data rather than collecting data to assess preconceived models, hypotheses, or theories in an inductive process.

In this study, the in-depth interview was used to learn how people construct their realities—how they view, define, and experience the world (Taylor & Bogdan, 1998, p. 101). The interviewer must be careful to allow the interviewee the opportunity to express his or her views without biasing the responses. The researcher must maintain a delicate balance: A relationship needs to develop between the interviewer and interviewee such that trust is established and yet the interviewer is able to get the needed questions answered.

As discussed in the following section, this element of training the students in the interview technique was more difficult than initially anticipated.

A Virtual Qualitative Lab

In summer 2008, the authors received a faculty initiative grant from Walden University to conduct a qualitative study on the current life experiences of people over 85 years (the oldest old). To have the opportunity to gain participants from across the United States and allow student participation they developed an online laboratory. E-college software provided the framework, which allowed for asynchronous discussions and the ability to post information, such as downloaded Microsoft Word files. The course, Independent Research, was an elective course available to doctoral students in the clinical, health, or general psychology specialties.

Walden University student listservs posted the availability of the project, with interested students completing applications. Twenty-nine individuals submitted applications, which were then sorted by area of the country. Priority was given to those students who had previously taken a qualitative research course and had previous experience in gerontology. This resulted in seven students being selected for and registered in the course. They were a geographically diverse group with individuals from Ohio, Missouri, Florida, Texas, Washington, Arizona, and California. The instructors lived in Montana and Minnesota.

Prior to the course starting, all students received an Olympus digital voice recorder, purchased through the grant. The instructors developed and tested interview questions in focus groups. The instructors also developed a poster to solicit for elderly participants, as well as consent forms, demographic questions, and scripts for contacting facilities, which were all sent to the students. They also submitted an application to Walden University's Institutional Review Board (IRB).

During the fall quarter, students received a list of readings on qualitative methods and gerontology, a copy of the grant application, and a copy of the IRB application (see detailed course assignments in the Appendix). Students also received a letter from the principal investigators (PIs) instructors, which described the project and business cards from the PIs, all of which they provided to potential facilities. Each participating
facility administrator signed a participation agreement, which the instructors forwarded to the IRB along with consent forms from each participant.

Students and instructors participated in three conference calls throughout the 12-week quarter (two were initially planned; a third was added, as discussed later). During the calls, the instructors provided additional training, answered questions, and clarified procedures. The instructors recorded and posted the calls in the classroom for later review, as needed.

**Interview Training**

Students received instructions on how to do qualitative interviewing and were required to practice interviewing a friend, then an older adult, recording and submitting the interviews via e-mail to the instructors. The digital recorders allowed the recordings to be directly downloaded and sent by e-mail. Both instructors reviewed and provided feedback on the recordings. Through the review of the recordings, it was found that students had difficulty with the interview process. As a result, an additional conference call was held in which all class members listened to one instructor interview the other, and students had the opportunity to critique the interview and ask questions. The students then interviewed an additional older individual as practice, the instructors reviewed this recording, and if satisfactory, the student could begin the actual research interviews. If the additional interview was not successful, the student received feedback and was required to do another practice interview.

**Research Procedures**

The 7 students contacted facilities, followed up with interested individuals, and completed six 3-hr interviews each, resulting in 42 interviews. The students submitted the recordings by e-mail, along with all paperwork, including the consent forms. To achieve a passing grade in the course, students had to submit these items by the end of the quarter.

**Course Assignments**

Along with the research recordings and paperwork just described, the students were required to post responses to two discussion questions per week concerning the research procedures and gerontology. They also were required to post a one-page journal entry per week on the research and course experience.

**Student Learning**

The course relied on two primary measures of student learning. First, the recordings directly evaluated their interview skills. The instructors reviewed the recordings, listening for the use of follow-up questions and maintaining a conversational style while asking the questions. Students received a critique of each practice recording, and needed instructor approval to continue.

Second, students submitted a one-page journal entry weekly. In the journal, they reflected on what they learned during the week, their feelings toward the course and the project, and any unmet needs. To examine these journal entries empirically, the researchers used Qualrus, a qualitative software program. They used entries from Weeks 1 and 2 (early part of the quarter), 6 and 7 (middle of the quarter), and 11 and 12 (end of quarter); they then downloaded the papers and coded for themes that reoccurred. The researchers randomly assigned students a letter in this paper to identify their responses yet preserve their anonymity.

In Week 1, 5 of the 7 students mentioned being nervous about conducting interviews and working with elderly individuals. Beginning as early as Week 2, students asked questions that could lead to future studies: “I wonder how or whether living through Hurricane Ike and the aftermath will influence the themes from the study” (student who lived in area of hurricane, Student B). “I also would like to ask at least one question up front and again at the end, to see how the project may have changed the participants’ perceptions” (Student G).

Other early comments also showed insights into critical thinking: “For instance, the VandenBos article discusses the concept of successful aging. Although there is no right or wrong answer, this challenged me to reflect upon my own past, present, and future life and how I would view this concept” (Week 2, Student F).

The reality of research was evident by Weeks 6 and 7, such as this comment about the difficulty of getting participants: “My main issue at this point is that the facilities have not yet contacted me back. It’s been over a week and I am a little worried” (Week 7, Student E).

By the end of the course, students showed an awareness of the social implications of the study: “One thing I have noticed about the caregivers at this facility is how they view the oldest-old. I have had contact with the directors or administrators . . . it seems that they forget that they too will one day get to be this age” (Week 11, Student C).
They also showed personal growth and development: “I have greatly increased my overall knowledge of the oldest-old, research that has been conducted on this topic, and feel a bit more comfortable and confident in moving forward with my dissertation topic” (Week 12, Student F). Student F (Week 11) captured the students’ experiences with “I feel a sense of sadness when I leave participants. I think this is because I wonder if I will ever see them again. I care about these folks well beyond just collecting data.”

Suggestions and Lessons Learned

At the end of the course, students indicated that they enjoyed giving their opinion of the interview questions and procedures, although it appeared to raise anxiety at the time. One difficulty with online instruction is that students tend to feel isolated and fear responsibilities that they cannot handle. Clear expectations and instructor availability via phone and e-mail in addition to their presence in the online course was essential to reduce student anxiety.

As previously described by Stadtlander (1998), personality issues occasionally caused a flurry of activity and concern between the students. At one point, a procedure was unclear and resulted in long online discussions between students and instructors, as they tried to clarify the item. Students also expressed that they would have liked to help with the analyses, which was not possible in the time involved.

Advanced preparation of materials allowed the course to run smoothly. Having two instructors was essential, as listening to and analyzing the large number of recordings would have been difficult for one person. Despite the difficulties related to training issues and student anxieties, the advantage of being able to collect a large amount of data in a geographically diverse population while also providing student research opportunities was well worth the effort. In the future, stronger requirements for participation in the study would be appropriate: Students must have a background in interviewing strategies, qualitative method, and strong interpersonal skills; the authors suggest using a telephone interview to screen applicants rather than just an application.

Summary and Implications

The advent of the Internet and graduate online learning has resulted in a gap in research mentoring that has been difficult to resolve. This article described a virtual qualitative research lab that both mentored students and gathered data from a large number of participants across the United States.

The project clearly demonstrates the feasibility of research mentoring online and suggests that distance-learning instructors, trained in the research process, are capable of mentoring and adequately conducting research. Such mentoring does require an appropriate research question and requires additional resources, such as conference call capability and digital recorders. Future instructors might wish to explore a quantitative research lab, with the opportunity for students to analyze the data collected.

References


Appendix

Course Assignment Summary

Week 1


Week 2

This week you will be reviewing the paperwork for the project, researching the mandated reporter laws for your state, and begin learning about gerontology.


Paperwork instructions for project

Week 3

This week you will continue learning about gerontology and begin learning about qualitative research.


Week 4

This week you will continue learning about qualitative research and learn about the Mini Mental Exam.


Mini Mental Status Exam: Download one or two of the MMSE from a Web search, and examine the issues that are identified through their use.

Week 5

This week you will continue learning about qualitative research.


Week 6

This week you will learn about the measures that we will be using.


Week 7

This week you will learn about some of the issues related to caregiving of older adults.

**Week 8**

This week you will look at research on gerontology, health, and illness.


**Week 9**

This week you will explore a review article.


**Week 10**

This week you are to find an article on the oldest old that relates in some way to our study. Write a brief summary in your discussion and offer your opinion, based on the info you have learned this quarter. Attach the article to your post.

**Week 11**

This week you are to find an article on the oldest old that relates in some way to our study. Write a brief summary in your discussion and offer your opinion, based on the info you have learned this quarter. Attach the article to your post.

**Required Activities**

There are a number of activities that you will be required to participate in. They are listed below.

**Week 1.** Carefully read the info sent to you concerning contacting facilities. Practice the script on several people. Report back on how you are feeling about it.

**Week 2.** Contact a nursing home and assisted living in your area and discuss project with director (materials). Post flyer in senior center. Research mandated reporter laws for your state. Conference call between all members of the course and instructors.

**Week 3.** Each student will interview a friend, and will need to submit the audiotape, an evaluation, and paperwork to the instructor by e-mail.

**Week 4.** Discussion of all problem areas online in class. Student will interview an elderly friend or relative for 1 hr. You will be assigned which day of interviewing you are to do, audiotape it and submit all materials for review by e-mail to the instructors by end of Week 5. Conference call between all members of the course and instructors. A practice interview will be done between the co-PIs on a conference call with the students listening; the students will then complete all paperwork and submit it for review by e-mail.

**Weeks 5–12.** Discussion of problem areas and approval to begin official interviews. **Do not begin interviews until you have been given official approval.** Begin interviews (each student will complete six 3-hr interviews of individuals over 85). Weekly readings and discussion on gerontology and methodology. All interviews will be recorded and the recordings and paperwork will be forwarded to the instructor as designated.

**Week 11.** Conference call between all members of the course and instructors.

**Week 12.** Final impressions on research experience and report of what you have completed this quarter.

**Notes**

1. We thank student researchers Mavis Christopher, Cathy Ferrarese, Kimberly Kohli, Cynthia Traylor, Miriam McCray, Tra Mayes, and Michael Warren. Our gratitude goes to Walden University for funding and supporting this project.

2. Send correspondence to Lee M. Stadtlander, 25 Morgan Creek Lane, Bozeman, MT 59718; e-mail: leann.stadtlander@waldenu.edu
Using a Cost-Benefit Analysis to Teach Ethics and Statistics
Jennifer DeNicolis Bragger and Mark A. Freeman
Teaching of Psychology 1999 26: 34
DOI: 10.1207/s15328023top2601_6

The online version of this article can be found at:
http://top.sagepub.com/content/26/1/34

Published by:
SAGE
http://www.sagepublications.com

On behalf of:

Society for the Teaching of Psychology

Additional services and information for Teaching of Psychology can be found at:

Email Alerts: http://top.sagepub.com/cgi/alerts
Subscriptions: http://top.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Jan 1, 1999
What is This?
METHODS AND TECHNIQUES

Using a Cost–Benefit Analysis to Teach Ethics and Statistics

Jennifer DeNicolis Bragger
Montclair State University

Mark A. Freeman
Temple University

We used Rosnow's (1990) method for quantifying the costs and benefits of research to teach undergraduate psychology students research ethics and the statistical concepts of central tendency and variability. Students rated the costs and benefits of well-known psychology experiments during the first class period and learned how to analyze and interpret their findings during the second lesson. We used class findings to demonstrate variability and consistency in ethical judgments of students. Students indicated they found the exercise useful for understanding research ethics and statistical concepts.

Teaching the American Psychological Association (APA; 1992) ethical principles of research to psychology undergraduates often involves the use of case studies (McMinn, 1988). In-depth discussion of specific research situations can demonstrate to students that there is often substantial variability in the judgment of ethical considerations in research but can leave students confused about the reasons behind this variability (Kimmel, 1991, 1996; Rosenthal & Rosnow, 1984; Rosnow, Rotheram-Borus, Ceci, Blanck, & Koocher, 1993). One way of explaining this variability to students, which in turn should improve critical thinking and be a vehicle for teaching them statistical concepts (Derry, Levin, & Schauble, 1995), is to quantify their ethical judgments of case studies and compare the judgments statistically.

The exercise described in this article is an adaptation of Rosnow's (1990; see also Strohmert & Skleder, 1992) method for rating the ethical costs and future benefits of case studies on a numerical scale and using these ratings to determine whether the research should be conducted. In the adaptation used here, the exercise served the dual purpose of teaching students research ethics and statistical measures such as means, standard deviations, and t tests. We incorporated some innovations that demonstrated to students how ethical judgments reflect elements of both variability and consistency. The exercise gave students a sense of the process by which institutional review boards (IRBs) evaluate research proposals, which should make students aware of the costs of conducting and of not conducting research (Rosenthal & Rosnow, 1984). In addition, the exercise demonstrated to students that (a) some variation in their ethical judgments of costs and benefits was predictable; (b) their ethical judgments, taken as a group, showed consistency; and (c) statistical concepts may be better understood when comparing one's own behavior to others.

Method

The participants were students in four undergraduate research methods classes in psychology who had completed at least one course in statistics. Generally, the relatively small number of students in a research methods class (approximately 24) makes the exercise appropriate for this course. We divided the discussion of the exercise into two class periods of 120 min each. During the first class, we discussed ethics in behavioral research and presented five prominent studies in social psychology. We used statistical measures of central tendency and spread to teach the concept of variability (i.e., within individuals and studies in a cost–benefit analysis) and to explain the decision process by which an IRB operates. During the second class we used students' ethics ratings to foster comprehension of systematic variability in statistical analysis.

Class 1

The first part of the exercise demonstrated three things to students: (a) the degree to which students tended to view research in terms of its ethical costs to research participants or future benefits to society; (b) the process by which the class, acting as a surrogate IRB, would decide whether each study should be conducted; and (c) the variability in cost–benefit assessments within individual student ratings of the five studies and within ratings of particular studies across individuals.

We began the first lesson with a review of an abbreviated version of the APA ethical guidelines as presented in an undergraduate text (Rosnow & Rosenthal, 1996). We instructed students to refer to the guidelines throughout the remainder of the exercise and then presented five research studies as cases for ethical review. The studies presented included Milgram's (1963) study of obedience, Rosenthal and Jacobson's (1968) study of expectancy effects in elementary schoolchildren, Darley and Latané's (1968) diffusion of responsibility study, Dutton and Aaron's (1974) investigation
of the effects of anxiety on sexual attraction, and Rosenhan's (1973) study of the classification of mental illness. We chose these studies because we believed their content and methodology would interest students, but other studies may be substituted. We informed students that their participation in this exercise was voluntary, and they could refuse to participate without penalty. There were no refusals.

We asked students to rate the costs and the benefits of each study separately on scales from 1 (lowest costs and benefits) to 100 (highest costs and benefits). Each student also generated an alphanumeric code, which revealed that person's gender but protected the person's identity. It would be possible to have students rate these studies for a homework assignment.

We asked a student to draw two grids, one for the cost assessment and one for the benefit assessment, on the board. We presented the studies in the rows and the raters in columns, so that a summary of studies and raters' identification numbers appeared in row and column margins.

We presented Rosnow and Rosenthal's (1997) decision-plane model, a graph that plots costs of research along one axis and the benefits along the other, to help students visualize the costs and benefits of the studies. We plotted cost and benefit ratings for each study on the decision plane, and we instructed students that research should be conducted when benefits were high and costs low and not conducted when costs were high and benefits low. We told students that researchers presume that IRBs make decisions in a similar manner, but they often examine costs in detail and pay little or no attention to potential benefits (Rosenthal & Rosnow, 1984).

The display of the ratings for each individual and each study allowed students to make an "eyeball comparison" of their cost and benefit evaluations for each study with those of their classmates. We discussed the variability of cost and benefit scores within their own ratings of the five studies and the variability of ratings across students within each study. Students used their calculators to compute the mean and range for each study and each rater and then entered the results in the grids. Students compared each of their cost and benefit ratings with the mean for that study and noted where their ratings fell along a continuum. Students compared their overall cost and benefit means with the means for the class and noted how costly or beneficial they perceived the studies to be relative to the class. We reviewed variability by computing standard deviations for men and women separately (see Table 1). The homework assignment was to compute standard deviations for each study for men and women by the next class meeting.

### Table 1. Means and Standard Deviations of Cost and Benefit Ratings for Men and Women for All Classes

<table>
<thead>
<tr>
<th></th>
<th>Men</th>
<th>Women</th>
<th></th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>n</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Benefits</td>
<td>69.37</td>
<td>11.09</td>
<td>30</td>
<td>60.11</td>
<td>15.85</td>
</tr>
<tr>
<td>Costs</td>
<td>53.30</td>
<td>11.51</td>
<td>30</td>
<td>57.01</td>
<td>15.81</td>
</tr>
</tbody>
</table>

### Class 2

The first class ended with students discussing variability in the assessment of the costs and benefits of research. We used the second class period to demonstrate the idea of consistent variability in ethical evaluations and in statistical analysis. Before class we redrew the grids on the board (it would also be possible to present the grids on an overhead to save class time). We compared the standard deviations for men and women for each of the studies. We pointed out to students that, although there is variability within the ratings by a single student or of a single study, the ratings by some students of certain studies were consistently higher or lower than others. The review of between-group and within-group variability also gave us an opportunity to review the meaning and interpretation of z scores and t tests. We discussed the use of ANOVAs in comparisons of more than two groups and explained that although a one-way ANOVA could be used to analyze these data, we used t tests because they provided more power and allowed initial focused comparison of the groups. The students computed a t test comparing the overall difference in ratings between men and women for cost and for benefit ratings. We pointed out the greater variability in the women's ratings compared with the men's. We then discussed the finding that men rated the studies higher on benefits, t(92) = 2.88, p = .002, two-tailed, and that although women appeared to rate the studies higher in costs, this difference was not statistically significant.

We discussed the idea of gender as a predictor of variation in ethical judgments and noted how the small sample size (for each class) affected statistical power and made it difficult to detect differences (Rosnow & Rosenthal, 1996). Interestingly, Kimmel (1991) made similar observations of differences between genders, which led us into the concept of replication. We explained to students that if we had predicted the direction of these findings based on Kimmel's research, then we could have reported a one-tailed p and thus obtained statistical significance for both t tests.

The class split into groups of three or four students and the groups computed t tests to determine if there were significant cost and benefit rating differences between men and women within each of the five studies. The t tests comparing cost and benefit ratings between men and women in individual classes were often not significant due to the small sample size, but when combining the ratings for all classes, women rated the Rosenhan (1973) study significantly higher in costs, t(92) = 2.19, p = .032, two-tailed, and men rated both the Milgram (1963) and the Dutton and Aaron (1974) studies significantly higher in benefits, t(92) = 2.51, p = .016, two-tailed, and t(92) = 2.20, p = .016, two-tailed, respectively. Figures 1 and 2 show that these gender differences, although not always significant, were consistent for all but Milgram's study of obedience, which both men and woman rated high on costs. The results of the tests prompted a discussion comparing statistical and practical significance.

After the discussion, we administered a questionnaire that assessed how students rated the exercise on a scale with anchors at 1 (low ratings) and 10 (high ratings). The students rated how useful the exercise was for teaching ethics in psychology (M = 6.93, SD = 2.51), how helpful the exercise was
in understanding the costs and benefits of research ($M = 7.17, SD = 2.22$), how strongly they would recommend using the exercise to teach ethics in psychology ($M = 6.62, SD = 2.37$), how useful the exercise was in helping them understand statistical measures of central tendency and spread ($M = 7.76, SD = 1.88$), and how strongly they would recommend using the exercise to help teach statistics to others ($M = 7.28, SD = 2.12$).

Thus, we conclude that the exercise was successful as an aid to understanding ethical principles and statistical analysis. We urge instructors to use such an exercise to generate more practical understanding of ethical considerations in the behavioral sciences.

References


Notes

1. We thank Ralph Rosnow for his comments on earlier drafts of this article.

2. Send correspondence to Jennifer DeNicolis Bragger, Montclair State University, 1 Normal Avenue, Upper Montclair, NJ 07043; e-mail: donb@webspan.net.

Teaching of Psychology
skills they reported to be the most valuable—skills such as critical thinking or writing. Eliminating these courses in an attempt to tailor curricula to career-bound students would effectively eliminate the very courses in which they developed valued job skills.

References


Notes

1. This research was supported by a Level 3 Grant for Assessment Activities administered by the Winona State University Office of Assessment.
2. Both authors contributed equally to this study.
3. Send correspondence to John C. Johanson, Department of Psychology, Winona State University, Winona, MN 55987–5838; e-mail: johanson@winona.edu.

Teaching Research Ethics: Illustrating the Nature of the Researcher–IRB Relationship

Thomas M. Brinthaupt
Middle Tennessee State University

I collected information on the effects of ethical concerns on research questions asked and methods used by psychological researchers. Faculty researchers described the advantages and disadvantages of ethical concerns in their specialized areas and provided case examples of the effects of institutional review board actions on their research questions and methods. I offer several suggestions for using this information to increase students’ ethical awareness and understanding in research methods courses.

Teaching ethics is an important part of research methods courses. Instructors typically need to help students understand what a researcher can and cannot do, ethically speaking. The instructor may have had experience with an institutional review board (IRB) and may need to guide students in the preparation of an IRB approval form for their research projects. In my experience, students often lack a clear understanding of the purposes of the IRB and the need to be ethically sensitive in their research. Most students are likely to be concerned mainly with the requirements of their instructors rather than of an IRB.

Authors have suggested various techniques for research methods instructors to address this lack of knowledge or concern about ethics (see Ware & Brewer, 1999). Among these techniques are attempts to get students to take the perspective of a research participant (Beins, 1993), to play the role of an IRB member who evaluates proposed research (Johnson & Corser, 1998), and to clarify their own values and ethics with regard to research possibilities (Rosnow, 1990). All of these approaches have the important goal of teaching students to realize, define, and understand what is ethical and what is unethical in psychological research. The hope is that students subsequently will show greater sensitivity and will make better decisions about the ethical aspects of their (and other) research.

There has been a great deal of research on IRBs over the past 25 years (for a review, see Landrum, 1999). Many commentators have called attention to the need for researcher awareness and education about ethical issues (e.g., Sieber, 1992). Some researchers have focused on the potential rewards and benefits of ethical sensitivity in one’s research. Blanck, Bellack, Rosnow, Rotheram-Borus, and Schooler (1992) illustrated how the principles of confidentiality, debriefing, and volunteer participants benefit the goals of researchers. Kallgren and Tauber (1996) advocated the use of IRB reviews for undergraduate research and showed that the review process enriches the educational experiences of student researchers. Other researchers have focused on the difficulties that researchers face when working with IRBs (e.g., Ceci, Peters, & Plotkin, 1985; Council, Smith, Kaster-Bundgaard, & Gladue, 1999; Gray & Cook, 1980; Liddle & Brazelton, 1996). For example, in Council et al.’s survey, nearly 50% of researchers believed that the IRB did not understand the nature of their research. On the other hand, a common IRB committee complaint was that researchers failed to follow standard guidelines for protocol submission.

Absent from both previous research and the techniques suggested for increasing ethical awareness in students has been the impact of IRBs on the questions that psychologists choose to ask and the methods they employ. Rosnow (1997) described the ethical guidelines that face psychologists as a continuously evolving “social contract” between science and society. As a means of educating students, another way instructors might supplement Rosnow’s important observations and recommendations is by providing information that reflects how societal concerns about ethics have an impact on researchers. In particular, how do ethical concerns positively and negatively affect research practice?

I propose an additional pedagogical approach for instructors in their efforts to enhance their students’ ethical sensitivity and understanding. This approach involves the use of information on faculty researchers’ attitudes toward and experiences with IRBs as well as actual examples of the impact of ethical concerns on the conduct of their research.
To get a clearer picture of researchers’ views on ethical issues and their experiences with IRBs, I surveyed 23 full-time psychology faculty from a large public state university and several smaller private colleges. Faculty recalled salient positive and negative experiences with IRB reviews of their research as they pertained to the conceptualization or “germination” of a research idea, the implementation of the research idea (methods, procedure), and working with the IRB. In addition to rating specific aspects of those experiences, they provided case examples of research that had been altered or revised due to ethical considerations. When my students and I discuss research ethics, I first present a summary of this information, followed by an in-depth discussion of it.

Among the major positives of ethical concerns reported by the researchers were effects on research participants (e.g., protection of participants’ well-being; increased trust and greater honesty in responding), effects on the researchers themselves (e.g., increased consideration of the perspective of potential participants; prevents the researchers from being the sole decider of what is ethically appropriate), and effects on the methods used (e.g., helps the researcher think clearly through the procedures or protocol). Among the major negatives were pragmatics of research (e.g., unnecessarily restrictive regulations), general negative effects (e.g., decreased external validity; methodology changes that compromised the value of the research), and negative effects on specific research topics (e.g., difficulties in observing actual behaviors of interest; limits to manipulations of strong emotions in memory research). Overall, faculty researchers reported that having to address ethical concerns had a generally negative effect on the development and conduct of research. Despite the drawbacks, researchers also recognized the positive impact of ethical concerns.

Table 1 presents some of the case examples I collected. During class discussions, I use these examples to draw attention to how faculty balance research goals with their own and others’ concerns about ethics. After we determine what the potential problem is, we discuss how the faculty researcher worked within IRB guidelines to solve that problem. We also discuss the researchers’ implicit and explicit assumptions conveyed by their case examples. For instance, with the second case example in Table 1, we discuss whether the IRB’s actions were “over-enthusiastic” or justified. I have found these discussions to be successful in highlighting the collaborative as opposed to the antagonistic relationship between researcher and IRB.

<table>
<thead>
<tr>
<th>Methodology changes that compromised the value of the research</th>
</tr>
</thead>
<tbody>
<tr>
<td>In one notable study negatively affected by IRB restrictions, we</td>
</tr>
<tr>
<td>only wanted to be able to link parents’ and children’s responses to our research protocol. Strict adherence to the IRB’s decision made it impossible to track changes in agreement over time, which was a major area of deficit in existing research. Thus, the study was of little benefit.</td>
</tr>
<tr>
<td>Idiosyncratic IRB definitions of “harm” to participants</td>
</tr>
<tr>
<td>IRB wanted to change questions on a marital adjustment questionnaire (the one most frequently used in the research literature); IRB thought that if you ask people “Have you ever thought about divorce?” then you’ll “put ideas in their heads.” IRB eventually allowed the use of the questionnaire, but study was delayed due to over-enthusiastic definition of “potential harm.”</td>
</tr>
<tr>
<td>Completely terminated the type of research I was drawn, trained, and would prefer to do</td>
</tr>
<tr>
<td>It changed [a long time] the whole tone of what I do; sent me to survey and questionnaire work which substantially decreased my motivation to do research in general.</td>
</tr>
<tr>
<td>Difficulties in accessing private information (e.g., abuse reports)</td>
</tr>
<tr>
<td>I planned to study parent/child interactions and compare groups by who has a record of physical abuse/maltreatment. Because parent consent for CPS [Child Protective Services] data was required, many parents refused to allow us to talk to CPS. Thus, our “groupings” were based on parents’ report from an abuse potential paper/pencil method rather than on actual official reports of abusive behavior.</td>
</tr>
<tr>
<td>Applied project difficulties</td>
</tr>
<tr>
<td>I have a lot of trouble determining what is appropriate on applied projects. For example, in our teamwork research, team members’ responses are anonymous and aggregated. However, the organization wants to use this information for management purposes. It can be used to help leaders work more effectively and/or get appropriate training. It may also make specific leaders look bad. Thus, the research and feedback doesn’t put the participants at risk, but it may have some consequences for others in the organization. Where does an organization’s legitimate rights to improve their company end and the supervisor’s personal rights begin?</td>
</tr>
<tr>
<td>Collection of sensitive materials</td>
</tr>
<tr>
<td>In one study, we omitted some items that were especially sensitive (which actually improved the materials), but we also took more steps in the collection phase (e.g., a lock box to drop questionnaires into) that added to the amount of work.</td>
</tr>
</tbody>
</table>

Note. IRB = Institutional Review Board.
view as an externally imposed requirement that all researchers must face. Clearly, conveying a balanced message makes more sense than conveying a negative view of the IRB as a common enemy or as creating “unwarranted meddling” in one’s research. Obviously, the message to students should not be that it is acceptable to ignore IRB procedures when they are under time pressure. Rather, instructors should point out to students that faculty can feel similar pressures when they conduct their research. The information I present to my students provides a good illustration of those pressures felt by faculty researchers. I have collected data showing that students who received and discussed this information show attitudes about ethical issues that are closer to faculty attitudes than students who did not receive it.

In conclusion, the major value of this approach is that it focuses directly on a neglected aspect of teaching and learning research ethics: the nature of the researcher-IRB relationship and how it can affect the research process at various stages. By focusing on this relationship, instructors have an additional tool for enhancing the ethical understanding of students. Even if students normally gain some understanding of the specific do’s and don’ts of ethical psychological research, incorporating this type of information may provide them with a more accurate appreciation of the bigger picture that researchers face. This information can be useful in beginning to complete that picture for students and in giving them a better appreciation of the dynamic nature of the researcher–society social contract Rosnow (1997) described.

References


Notes

1. Preparation of this article was supported by a Middle Tennessee State University Non-Instructional Assignment grant.
2. I thank Will Langston, John Pennington, Terry Whiteside, and the editor and reviewers for their comments on earlier drafts of the article.
3. Send correspondence and requests for further details on the faculty respondents, survey instruments, procedure, and evaluation data to Tom Brinthaupt, P.O. Box X034, Department of Psychology, Middle Tennessee State University, Murfreesboro, TN 37132; e-mail: tbrinthta@mtsu.edu.

Textbook Selection: Balance Between the Pedagogy, the Publisher, and the Student

R. Eric Landrum
LuAnne Hormel
Boise State University

We report a national survey of introductory psychology instructors about how pedagogical aids and publisher practices impact textbook selection and student learning. Respondents evaluated the importance of 79 attributes on 2 separate criteria: textbook selection and facilitating student learning. Some criteria received high ratings on both dimensions (e.g., textbook accuracy, readability/writing quality, use of examples on both textbook selection and student learning), whereas others received high ratings on only 1 dimension (i.e., ancillary materials and test item files for the textbook selection process). Experience in teaching the course also influences the relative importance of textbook selection criteria.

“One of the most important decisions an instructor makes is the selection of a textbook” (Chatman & Goetz, 1985, p. 150). Although not all instructors would agree with the previous statement, this decision is both an important one and also a difficult and daunting task. Griggs, Jackson, Christoper, and Marek (1999) reported that there are 37 full-length introductory psychology textbooks available; there are numerous other “basic” or “briefer” editions available. For instructors a critical question becomes “Which textbook do I choose?” Our interest was to understand (a) the process of selecting a textbook and (b) the instructor’s perceptions of textbook’s role in facilitating student learning.

Most of the literature on textbook selection is advisory in nature. These suggestions offered different methods used (Chatman & Goetz, 1985), including a textbook selection committee (Britton, Woodward, & Binkley, 1993), student
Introducing Diverse Perspectives Into Research Methods Classes

Catherine Borshuk
Indiana University South Bend

Instructors of undergraduate research methods can introduce diverse perspectives into their courses through expanding learning units on research ethics to include extensive discussions on the responsibilities of the researcher. I provide suggestions for teaching strategies that promote multiculturalism while avoiding a deficit research perspective.

Content courses in psychology represent a natural opportunity to challenge bias, introduce diverse perspectives, and integrate cultural understandings while teaching principles of psychological knowledge (Gloria, Rieckmann, & Rush, 2000; Goldstein, 1997; Kowalski, 2000; Simoni, 1996). In this article, however, I address the need and the opportunity for psychology instructors to similarly incorporate discussions of diversity into undergraduate research methods classes. I have developed strategies to integrate diverse perspectives in both experimental and nonexperimental research classes, largely by revising the way I teach ethical principles for research in psychology (American Psychological Association, 2002). Including such perspectives is especially important because many psychology graduates are likely to work with disadvantaged communities and in applied settings where diversity arguably matters more than in research settings.

Students are often unfamiliar with the idea that research can be harmful to people or communities. Therefore, I introduce the topic of research ethics by describing studies, such as the Tuskegee Syphilis Study (Jones, 1981), that illustrate the dangers of the attitudes of majority-group researchers toward minority-group research participants and of the consequences of a lack of informed consent. My students hold small group discussions about whether the social beliefs of the researchers reflected racism and colonialism (Tuhiwai Smith, 2002, effectively drew the analogy of researcher to colonizer). Once they understand the way that research can cause harm, they are able to discuss important questions, such as these: Who pays the price for research, and who benefits from it? In what ways are social relationships reproduced in the research relationship? I also use this case to illustrate long-term consequences of such research. For example, recent survey findings reveal that over half of African Americans believe that a cure for HIV/AIDS is being deliberately withheld from sufferers (Bogart & Thorburn, 2005), perhaps because of lingering mistrust of health professionals and researchers.

I use a similar strategy of presenting a historical example, followed by small group discussion, to illustrate the way that research practices have affected indigenous peoples. Many Native communities no longer allow outside researchers to collect data (Kral, Burkhardt, & Kidd, 2002). For instance, Cree leaders in northern Quebec declared a moratorium on psychological research due to “the researchers’ refusal to accept Cree authority and to the little perceived benefits of this research for the community” (Darou, Kurtnes, & Hum, 2000, p. 44). An Alaskan native community, harmed by the negative publicity generated by an earlier study on alcoholism, instituted a tribal ban on further inquiries into this important social problem (Norton & Manson, 1996). These examples offer me the chance to introduce the idea of deprivation or deficit research (Parham, White, & Ajamu, 2000), so called because it is based on the presumption that a minority community is of interest only to the degree that it poses research problems for majority-group researchers. Deficit research socially constructs communities as being naturally or culturally deficient relative to the standards of a European, White middle class. Researchers working from this model may view particular groups as embodiments of social problems and view social problems as puzzles that only educated outsiders are able to solve (Tuhiwai Smith, 2002).

Recently, more culturally appropriate research has challenged such notions by investigating the strengths and resilience of disempowered communities (e.g., Chataway, 2002). Educators and students can think more critically about the nature of deficit versus difference by being exposed to culturally different explanations for behavior that may not be found in standard psychology textbooks. In a nonexperimental research methods class, I introduce as one resolution to the co-
nundrum of the deficit perspective a research method that is gathering interest and adherents in applied and field settings, participatory action research (PAR).

PAR emerged in the 1980s in response to complaints that social scientific researchers were the sole beneficiaries of research. Because PAR primarily uses qualitative research methods, it may be less relevant to a class on experimental research methods, although its aims of overcoming the stigma associated with deficit research in minority communities is applicable in any discussion of responsible research. In PAR, the community members become involved as active research collaborators and contribute throughout the research process, from identifying the research question to making decisions about where to disseminate findings and how to implement action strategies (Maguire, 1987). PAR is a method that “makes an explicit commitment to working with members of communities that have traditionally been exploited and oppressed, in a united effort to bring about fundamental social change” (Brydon-Miller, 1997, p. 658). Psychologists are using PAR to examine the so-called achievement gap in educational attainment of African American youth (renamed the opportunity gap by researcher-participants; Fine & Torre, 2004) and in studies of strength and capacity in Canadian Aboriginal communities (Chataway, 2002). Psychology students who will one day work with disadvantaged communities need to be exposed to a critical understanding of deficit research and to alternative methods of investigating social problems.

Expanding the discussion of ethics in research methods classes to encompass the idea of responsible research allows the class to engage in discussions or demonstrations that encourage an understanding of psychological research and promote critical thinking about topics such as race, gender, and sexuality. In Table 1, I present a list of questions to promote group discussion, to structure assignments, or to develop exercises.

### Table 1. Suggested Class Discussion Questions and Exercises

<table>
<thead>
<tr>
<th>Questions for Discussion and Exercises</th>
<th>Description of Discussions and Exercises</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whose interests have been overlooked in psychological research?</td>
<td>Psychological research historically neglected women or assumed that certain constructs (e.g., achievement) were irrelevant to them (Caplan &amp; Caplan, 1994). Students can pursue this thread by examining groups that continue to remain underrepresented in psychological research, such as the poor, immigrants, and people of color (Graham, 1992).</td>
</tr>
<tr>
<td>How can we help to overcome the mistrust that certain communities have of outside researchers?</td>
<td>Students may compare the results of traditional social science research findings on indigenous peoples to findings that resulted from participatory action research projects to discover the differences in capacity approaches versus deficit approaches to research.</td>
</tr>
<tr>
<td>What is the best way to collect demographic data on participants? Is a researcher better off asking about subjective identity or objective identification (Reid, 2002)? What is the difference between a statistical norm and being socially normal?</td>
<td>I ask students to qualitatively describe themselves and then work in small groups to try to map their self-descriptions onto a demographic checklist they create together. I follow with large group discussions about the experience of having personal details collected in various ways and about how researchers may more sensitively gather these data. A body of research on mate selection, attractiveness, and intimate relationships, for example, presupposes a heterosexual norm that could further marginalize same-sex or nontraditional partnerships.</td>
</tr>
<tr>
<td>What do research and evaluation mean to different communities, and how may different understandings lead to problems in cross-cultural research settings?</td>
<td>To help students understand how different communities experience research issues, instructors may take advantage of guest speakers who can promote cross-cultural understanding. Such guests, if available, can promote a respect for nonacademic ways of producing knowledge and can provide understanding about how research is seen from the eyes of participants.</td>
</tr>
<tr>
<td>What does the multicultural psychology community recommend to researchers working with diverse populations?</td>
<td>The Council of National Psychological Associations for the Advancement of Ethnic Minority Interests (2000) prepared guidelines for researchers working in Asian American, African American, Hispanic, and American Indian communities. The guidelines focus specifically on cultural sensitivity and cross-cultural knowledge, reinforcing how essential diversity is for researchers.</td>
</tr>
</tbody>
</table>

### Conclusion

It is possible to infuse diversity into undergraduate research methods courses in addition to psychology courses that focus primarily on disciplinary content. Instructors can teach research ethics in a way that demonstrates the unique responsibility of the researcher to both the discipline and to the community being studied. Along with learning about sampling and the finer points of experimental design, students need to confront the ways that research can, in fact, be harmful to individuals or populations. Using such cautionary examples helps students to learn the importance of respect for the human individual, the goal of doing no harm in the research process, and the responsibility to treat participants with justice and to protect them from exploitation, principles that are articulated in The Belmont Report (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979). Requiring students to pay spe-
cial attention to diversity issues when designing studies or research proposals signals that the promotion of tolerance extends beyond mere lip service in psychological science and practice and that examining assumptions about other groups may lead to research that is relevant and consequential to individuals beyond the academy.

References


Note

Send correspondence concerning this article to Catherine Borshuk, Department of Psychology, Indiana University South Bend, 1700 Mishawaka Avenue, South Bend, IN 46634–7111; e-mail: cborshuk@iusb.edu.

Interviews With Primary-Caregiving Fathers Via E-Mail

April Schwarzmueller
Human Development Discipline
Eckerd College

Students in a child development course interviewed primary-caregiving fathers (PCFs) via e-mail and wrote formal papers discussing their interview data with reference to relevant empirical literature. Analyses showed the project provided students a valuable opportunity to learn about challenges PCFs face, gender stereotypes about parenting, and fathers’ influence on children’s understanding of gender roles.

Students of developmental psychology should learn about nontraditional as well as traditional patterns of childrearing (Mazur, 1999). One nontraditional parenting arrangement involves primary-caregiving fathers (PCFs); there are currently 98,000 PCFs in the United States (Fields, 2004). Research indicates undergraduates hold negative perceptions of this arrangement (Riggs, 1997; Wentworth & Chell, 2001). The goals of the project were for students to (a) consider challenges PCFs face, (b) explore gender stereotypes about parenting, and (c) think about the role fathers play in children’s understanding of gender roles.

Although PCF support networks exist from which professors may find guest speakers, not all cities have such groups. To overcome this limitation, this project enabled students to interview PCFs by e-mail. Although literature exists on various pedagogical uses of Internet technology, there is little published research on the use of e-mail interviews in college teaching (for an example of such research, see Gilbride, 2000).
Undergraduate Research Experiences in Psychology: A National Study of Courses and Curricula

Baron Perlman and Lee I. McCann
University of Wisconsin Oshkosh

We surveyed departments nationally to better understand the extent of scientific opportunities and experiences for undergraduate psychology students. Results showed intradepartmental variability, but overall students can expect 7 courses that offer research experiences in the typical psychology curriculum. Nonetheless, research is often not the primary course focus; some students must wait until their junior year to take such a course, and most such courses are elective, not required. We discuss implications for departments’ curricula and the goals and outcomes for undergraduate education as well as future research directions.

"The fundamental goal of education in psychology, from which all the others follow, is to teach students to think as scientists about behavior" (Brewer et al., 1993, p. 169). Not surprisingly, faculty ranked thinking of and teaching psychology as a science first in importance as an expectation for the undergraduate psychology major (McGovern & Hawks, 1986). How best to meet this goal? Psychology faculty use various methods to teach psychology's scientific procedures, including lecture, demonstration, discussion, and collaborative learning. However, "those who have studied the learning of science have concluded that students learn best if they are engaged in active learning" (Committee on Undergraduate Science Education, 1997, p. 4).

McKeachie and Milholland (1961) listed four goals for a good laboratory (research) experience: (a) appreciation of research and scientific method, (b) knowledge of methodology, (c) development of laboratory [research] skills, and (d) critical thinking and experience with the process of knowledge generation or discovery (pp. 74–75). Similarly, Kulik (1973) emphasized "the value of research experiences where students produce a finished product or can see a behavioral effect" (p. 208). McGovern, Furumoto, Halpern, Kimble, and McKeachie (1991) underscored the importance of "the skills to use experimental methods, statistics, and qualitative methods" (p. 601) and noted that laboratory work can develop these skills. Brewer et al. (1993) emphasized "opportunities for them [students] to integrate research, theory, and practice ... in the laboratory and in real-life situations" (p. 171). The importance of research experience also appeared in the quality principles for undergraduate education in psychology (McGovern & Reich, 1996) that encouraged "(a) multiple opportunities for students to be active and collaborative learners and (b) research projects to help students learn the science of psychology" (p. 255). Finally, the discipline’s learning goals and outcomes for majors (Halonen et al., 2003) state that "students can think as scientists ... and have developed skills and values that reflect psychology as both a science and an applied field" (p. 3).

Berthold, Hakala, and Goff (2003) emphasized the importance of students developing an appreciation of the laboratory (scientific) nature of the discipline, and Elmes (2002) identified important outcomes of this positive valuation as a result of doing scientific work: (a) a higher level of student engagement, (b) enhanced intellectual achievement in psychology, (c) better research performance, and (d) highly desirable skills for doctoral programs and employers. Wolfe, Reynolds, and Krantz (2002) suggested that the outcomes of undergraduate laboratory experiences include (a) developing writing, speaking, and reading skills; (b) sampling "the complex and messy 'real world'" (p. 7); (c) developing reflective judgment; (d) learning that not understanding can be exciting; (e) making students self-learners; (f) helping students understand psychology's science and methodology; and (g) assisting students in deciding if a major or further education in psychology is desirable. Berthold et al. suggested that a laboratory in introductory psychology should be the norm, not the exception, whereas Wolfe et al. echoed past conclusions by arguing for research opportunities beyond general laboratory classes.

All of these sources have agreed that research experience is a crucial component of a good undergraduate psychology education. Ideally, this agreement would be apparent in the structure of existing psychology curricula that would in turn build on the base of scientific training acquired from and in concert with general education courses. However, psychology students are not getting active scientific experiences in their general education and nonpsychology requirements as often as faculty may believe (Perlman & McCann, 1993), increasing the importance of research experiences within the discipline’s curricula and courses.

Experimental psychology is the most frequently listed course in department offerings likely to require student research (Scheier & Rogers, 1985). Yet as far back as The Cornell Conference (Buxton et al., 1952), there was dissatisfaction with the then traditional experimental psychology course with its weekly laboratory exercise that often failed to provide students an overall view of science or to develop positive student attitudes toward scientific methods in psychology. Nonetheless, Buxton et al. believed that labora-
tory experience was the best way to achieve these outcomes if courses (a) focused more on substantive problems than pure methodology, (b) included projects extending beyond a week, and (c) allowed greater student participation in planning and evaluating the experiments (greater student responsibility over their scientific work). McKeachie and Milholland (1961) concurred on the importance of laboratory work and noted that an ideal curriculum has a laboratory sequence ending with the student being responsible for all steps in conducting an experiment. However, they pointed out that laboratory work as part of a short (one semester) introductory course—and we extrapolate to other one-semester courses—creates pedagogical difficulties because (a) content must be adequately covered, (b) it may be difficult to coordinate laboratory work and content, (c) range and complexity of experiments may be limited, and (d) laboratory work may be subordinated to other course requirements.

The traditional experimental psychology course notwithstanding, most undergraduate psychology curricula lack the numerous laboratory courses and hours typically found in the natural sciences. For example, chemistry emphasizes the liberal arts in undergraduate education as does psychology, yet, in chemistry (Committee on Professional Training, 2003) guidelines defining high quality undergraduate programs that departments must meet for approval or certification require a core, consisting in part, of 7 semester credit hours (300 to 350 contact hours) of laboratory instruction distributed across subdisciplines. All such courses must be taught annually, and advanced courses bring total laboratory contact hours to 500. A major in physical chemistry also requires laboratory work in physics, and the chemistry minor includes 200 laboratory contact hours in at least two different areas.

More recent curriculum reports in psychology revealed that laboratory experiences were not integral to most required core courses and were limited in number, even in elective courses (Messer, Griggs, & Jackson, 1999; Perlman & McCann, 1999b). Scheier and Rogers (1985) noted that departments in 4-year schools were most likely to offer a laboratory (research) experience in the experimental psychology course (42%), with labs also found in physiological (16%), statistics (16%), social/personality (7%), introductory (5%), and developmental (5%). They reported that 7.1% of all courses in 4-year schools had a “significant” empirical research requirement. In more recent research, 23% of experimental courses had laboratory activities; physiological or biological psychology courses, 17%; learning courses, 14%; statistics, 12%; cognitive, 8%; social, 7%; and introductory, 5% (Perlman & McCann, 1999b).

Cooney and Griffith (1994) reported that 72% of psychology departments in 4-year schools required a laboratory or structured research experience, 43% required an individual research project, and 3% required research assistance to faculty. Their data highlighted the variety of scientific experiential opportunities available to students, including research assistance to faculty, structured curricular offerings such as labs, and individual research projects. Terry (1996) reported that 86% of departments offered research opportunities and 48% required research for the major, whereas another 14% sometimes required it. In addition, 27% of these departments required an undergraduate thesis. About half of reporting psychology department chairs at primarily undergraduate institutions indicated research supervision was not computed in faculty’s teaching loads. We have no comparable natural science data, but in chemistry (Committee on Professional Training, 2003) the maximum faculty teaching load in an approved program of 15 contact hours per week includes actual time in classroom and laboratories, not total number of course credits taught.

In none of these studies was the nature of these psychology research experiences investigated. What do students actually do when they practice psychological research? What types of experience do they gain, and how many participate? Such data are critical to the discipline’s understanding and meaningful discussion of how it teaches the science of psychology.

In this study we investigated undergraduate research experiences in psychology and how they fit into departments’ curricula. We defined students doing science broadly: controlled laboratory experiments, computer simulations, field observations, or analyzing and interpreting archival data. Our definition included the traditional experimental manipulation of causal variables, the psychometric study of individual differences and variation (correlational procedures), and descriptive “questionnaire” efforts.

We examined an array of matters pertaining to students’ research opportunities. For example, what percentage of undergraduate scientific work involves “cookbook” experiments (often perceived as dull and routine; Kulik, 1973) or computer simulations vis-à-vis replicated or original research actively conducted by students? Do faculty receive teaching credit for their work? We sought information on these and other questions about undergraduate research experiences in psychology.

Method

Participants

We studied the curriculum of 500 psychology “departments” (some 2-year schools have no psychology department but do offer psychology courses) in four institutional types using the Carnegie Foundation for the Advancement of Teaching taxonomy (Carnegie Foundation, 2000): (a) Doctoral/Research Universities—Extensive and Intensive, n = 133; (b) Master’s (Comprehensive) Colleges and Universities I and II, n = 133; (c) Baccalaureate Colleges Liberal Arts and General, n = 134; and (d) Associate’s Colleges (2-year institutions), n = 100.

Instrument

The questionnaire measured different dimensions of student research activity. We modified the questionnaire for 2-year institutions, deleting questions about number of psychology majors, the “thesis” or major senior experience, and students assisting graduate students with research.

Departmental and institutional context. We asked about (a) a department’s highest degree offered, (b) number
of psychology majors, (c) number of full-time equivalent faculty, (d) calendar system, and (e) public or private status of the institution.

How the curriculum structures research opportunities. We asked about (a) whether courses with research are required for majors or “typical” psychology minors; (b) the number of course sections offered per academic year in which students do research and enrollment for all sections; (c) the first course in the curriculum with research opportunities and its level (e.g., freshman, sophomore, junior, senior); (d) requirement of an undergraduate “thesis” or a substantive project for majors; (e) availability of five types of research opportunities to psychology majors and whether each was elective or required (classes with structured research experiences, e.g., labs; classes with research experiences in nonlaboratory settings, e.g., field research, observational studies; individual research projects designed/conducted by students; students working as research “assistants” to, or collaboratively with, faculty or graduate students); and (f) if faculty receive teaching credit for courses with research experiences.

Nature of research opportunities. We asked for information pertaining to (a) course name, number, and number of credits and (b) if research was the course’s primary or secondary goal, or whether both content and research were equally balanced. Respondents could check all research experiences in each course that applied: (a) instructor generated ideas for in-class exercises that result in data; (b) instructor generated ideas and design for larger projects; (c) computer simulations; (d) students analyzed preexisting data; (e) at least one research project extended beyond 1 to 2 weeks; (f) used an idea bank as an “experiment starter” with students refining the ideas and being fully responsible for the research; (g) data gathered on campus (e.g., labs, observational research); (h) data gathered off campus (e.g., field research, from organizations or files); (i) collaborative/interdisciplinary research opportunities with other departments; (j) students worked individually, in groups, or both; (k) data gathered from animals or humans (self, classmates, other participants); and (m) data shared via a paper (turned in to instructor), a class poster session, oral presentations in the course or community, or by other means.

Future research opportunities. We asked if over the next 5 years the number of research opportunities for psychology students at the institution would probably decrease, stay the same, or increase.

Procedure

We read psychology department information and course listings in institutional catalogs on the continuously updated online college catalog collection “CollegeSource” (Career Guidance Foundation, 1997). We sampled every 5th institution on the list for all but Associate Colleges (every 10th) until we completed data collection. We omitted catalog entries lacking course listings (Doctorate/Research, n = 14; Master’s, n = 10; Baccalaureate, n = 8, 2-year, n = 44). We identified courses with research opportunities and pre-entered data obtained from each department’s catalog descriptions on their questionnaire, noting whether the institution was public or private.

We asked chairs or a more knowledgeable colleague to approve or correct these data and to complete the remainder of the questionnaire and offer comments. Each questionnaire contained a page for each course identified by the catalog reader as having research opportunities plus two blank pages for courses we failed to identify or that were too new to be included in the catalog. Two weeks later we sent a follow-up letter, thanking those faculty who had responded and urging others to do so.

Catalog Versus Respondent Data and Level of the Data

We gathered catalog data for all 500 departments on number of required courses with research, content area, types of research opportunities for undergraduate majors, curricular dimensions of the research experience, research experiences, level and title (content area) of first research courses, and a required thesis. Respondent data differed from catalog data by 7% to 32% (catalog data collection errors of omission and commission) indicating that even current catalog material was sometimes incomplete or unclear (e.g., more courses required than we identified, courses with research opportunities earlier in the curriculum than we could find). Because of these differences, we lack confidence in the catalog data except for number and area of research courses (Table 1) and present only respondents’ data in all other areas of study.

We coded course data into seven general content areas (see footnote to Table 1) because of the large number of individual courses (N = 1,395), many with different titles and foci. The levels of analyses are institutional/department type (associate, bachelor, master’s, and doctoral) and course content area, not specific courses per se. We report number of courses in each content area but do not know how many different specific courses are tallied (i.e., statistics, research design, introduction to research methods are all coded in the research/experimental area).

Results

Type of Institution and Highest Departmental Degree

We received 203 responses (41%), 22 for associate colleges (22%), 59 for baccalaureate (44%), 58 for master’s (44%), and 64 for doctorate (49%). One hundred and eighty-one of 400 departments in 4-year, master’s, or doctoral institutions responded (45%). Of the 64 departments in doctoral universities, 59 offered a doctorate in psychology, 3 a master’s and 2 a bachelor’s degree. Of 58 master’s institutions, 31 departments offered a MA or MS, the rest a BA or BS degree.

Public–Private

Departmental institutions were 54% public (n = 109) and 46% private (n = 94). Two-year institutions were overwhelmingly public (91%, n = 20), bachelor’s primarily private (86%, n = 51), master’s more evenly split (59% public, n
Table 1. Average Number of Courses With Research Opportunities for Undergraduates in Psychology in Seven Content Areas

<table>
<thead>
<tr>
<th>Course Area</th>
<th>RQ</th>
<th>C</th>
<th>RQ</th>
<th>C</th>
<th>RQ</th>
<th>C</th>
<th>RQ</th>
<th>C</th>
<th>RQ</th>
<th>C</th>
<th>RQ</th>
<th>C</th>
<th>RQ</th>
<th>C</th>
<th>RQ</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>0.41</td>
<td>0.50</td>
<td>0.12</td>
<td>0.33</td>
<td>0.19</td>
<td>0.39</td>
<td>0.12</td>
<td>0.39</td>
<td>0.31</td>
<td>0.82</td>
<td>0.19</td>
<td>0.59</td>
<td>0.16</td>
<td>0.41</td>
<td>0.14</td>
<td>0.37</td>
</tr>
<tr>
<td>RQ</td>
<td>0.45</td>
<td>1.0</td>
<td>0.20</td>
<td>0.59</td>
<td>2.3</td>
<td>1.4</td>
<td>1.7</td>
<td>1.3</td>
<td>2.0</td>
<td>0.95</td>
<td>1.9</td>
<td>1.3</td>
<td>1.7</td>
<td>0.99</td>
<td>1.4</td>
<td>0.99</td>
</tr>
<tr>
<td>Methodology/experimental</td>
<td>0.68</td>
<td>0.71</td>
<td>0.23</td>
<td>0.51</td>
<td>0.88</td>
<td>0.17</td>
<td>0.90</td>
<td>1.8</td>
<td>0.47</td>
<td>0.80</td>
<td>0.47</td>
<td>0.84</td>
<td>0.89</td>
<td>1.4</td>
<td>0.83</td>
<td>1.3</td>
</tr>
<tr>
<td>Social/development</td>
<td>0.09</td>
<td>0.29</td>
<td>0.07</td>
<td>0.26</td>
<td>1.9</td>
<td>2.2</td>
<td>1.5</td>
<td>1.9</td>
<td>0.85</td>
<td>1.5</td>
<td>0.82</td>
<td>1.4</td>
<td>1.7</td>
<td>2.3</td>
<td>2.0</td>
<td>2.6</td>
</tr>
<tr>
<td>Natural science</td>
<td>0.09</td>
<td>0.29</td>
<td>0.02</td>
<td>0.14</td>
<td>0.24</td>
<td>0.54</td>
<td>0.25</td>
<td>0.56</td>
<td>0.16</td>
<td>0.41</td>
<td>0.13</td>
<td>0.40</td>
<td>0.23</td>
<td>0.50</td>
<td>0.35</td>
<td>0.60</td>
</tr>
<tr>
<td>Biological</td>
<td>0.13</td>
<td>0.35</td>
<td>0.10</td>
<td>0.39</td>
<td>0.49</td>
<td>0.70</td>
<td>0.47</td>
<td>0.65</td>
<td>0.78</td>
<td>0.80</td>
<td>1.0</td>
<td>1.0</td>
<td>0.66</td>
<td>0.82</td>
<td>0.80</td>
<td>0.90</td>
</tr>
<tr>
<td>Vocational</td>
<td>0.18</td>
<td>0.50</td>
<td>0.30</td>
<td>0.67</td>
<td>1.9</td>
<td>1.4</td>
<td>1.7</td>
<td>1.3</td>
<td>1.8</td>
<td>1.4</td>
<td>2.0</td>
<td>1.7</td>
<td>2.7</td>
<td>1.4</td>
<td>2.6</td>
<td>1.6</td>
</tr>
<tr>
<td>Total</td>
<td>2.1</td>
<td>1.8</td>
<td>1.0</td>
<td>1.4</td>
<td>8.0</td>
<td>5.0</td>
<td>6.6</td>
<td>4.6</td>
<td>6.3</td>
<td>3.2</td>
<td>6.5</td>
<td>3.7</td>
<td>8.0</td>
<td>4.7</td>
<td>8.1</td>
<td>4.6</td>
</tr>
</tbody>
</table>

Note. There were two sets of data: returned questionnaires, associate, n = 19; bachelor, n = 59; master, n = 58; doctoral, n = 64. We read 100 catalogs for 2-year schools, bachelor’s = 134, master’s = 133, doctoral = 133. RQ = returned questionnaire data; C = catalog data.

For example: statistics, experimental, research methods, tests and measurement. For example: social, child, developmental, human growth/development, adolescent, adult development, exceptional child. For example: industrial/organizational, counseling, field experience, educational. For example: independent study, special topics, seminar in psychology, research participation.

Number of Psychology Majors and Faculty

The 157 4-year or above departments averaged 317 majors (SD = 373, Mdn = 170). Four-year schools averaged 96 majors (SD = 64, Mdn = 80) with 5 departments having 30 or fewer and 5 having 200 or more. Master’s institutions averaged 207 majors (SD = 180, Mdn = 150) with 4 having 40 or fewer and 6 having 400 or more (1 reported 915), whereas doctoral institutions averaged 620 majors (SD = 461, Mdn = 500) with 6 having 150 or fewer (40 was the lowest) and 6 having 1,200 or more (1 reported 2,600). There were significantly more majors in public institutions (M = 543, SD = 446) than private (M = 185, SD = 159), t(102) = 4.50, p < .0001 (two-tailed), r = .41, and significantly more faculty (public: M = 20.6, SD = 12.4; private: M = 11.4, SD = 11.6), t(102) = 3.68, p < .0001 (two-tailed), r = .34.

Departments in both 4-year and master’s institutions had small faculties, especially given their numbers of majors, suggesting that there were relatively few faculty available to teach research skills in these institutions. The 2-year schools (n = 21) averaged 2.7 faculty teaching psychology (SD = 1.3, Mdn = 2.5). For other than 2-year schools (157 departments) the average number of faculty was 13.7 (SD = 11.7, Mdn = 9.0). Four-year schools averaged 6 faculty (SD = 2.9, Mdn = 6), master’s 10.3 faculty (SD = 8.8, Mdn = 8), and doctoral 25 faculty (SD = 11.6, Mdn = 23).

For 152 non–2-year schools we computed the average number of majors per faculty member (M = 21.3, SD = 11.4, Mdn = 20). The differences among departments in different types of institutions were not striking, signifying that availability of faculty to majors depended more on job responsibilities, department goals and mission, and faculty interests than numbers of majors or faculty. We computed an ANOVA and found a significant effect of institutional type on number of majors per faculty, F(2, 149) = 5.97, p < .003 (η² = .07). Post hoc analyses using Fisher’s LSD revealed bachelors’ institutions had the lowest ratio, (M = 17.1, SD = 10.3, Mdn = 15.8), compared with master’s (M = 22.3, SD = 11.8, Mdn = 20) and doctorate (M = 24.4, SD = 11.2, Mdn = 23.6) schools, which did not differ from each other. Master’s and doctoral public institutions averaged 25.3 majors per faculty (SD = 10.9), significantly higher than private institutions (M = 19.1, SD = 11.8), t(98) = 2.6, p < .012 (two-tailed), r = .25.

Research Opportunities: The Typical Curriculum

How many research opportunities awaited students, in what content areas (we categorized courses into seven areas, see footnote to Table 1), and were there differences among types of schools? Table 1 displays these data for both the catalog sample of 300 and the 203 responding departments. The data are comparable except for the 2-year schools (low response rate), possibly because the questionnaire may have elicited responses primarily from 2-year schools that offered research experiences. These data may, therefore, over-report research availability in such institutions.

Respondents reported an average of 6.9 courses with research opportunities in a typical department. There were considerable departmental differences in the total number of courses with research experience and the curricular areas in which they were offered. Large differences existed among the numbers of courses in various content areas, but not between types of schools. The methodology/experimental (M = 1.8
courses, \( SD = 1.3 \), natural science \( M = 1.4 \) courses, \( SD = 2.0 \), and advanced/special topics areas \( M = 2.0 \) courses, \( SD = 1.5 \) accounted for most of these courses.

The average ratio of number of research courses per faculty (for questionnaire data, \( n = 157 \)) for non–2-year departments was 0.94 \( (SD = .78, \text{Mdn} = 0.7) \), almost one course per faculty member. We computed an ANOVA and found a significant effect of institutional type on number of courses with research opportunities per faculty, \( F(2, 154) = 34.2, \ p < .001 \) \( (\eta^2 = .31) \). Post hoc analyses using Fisher’s LSD revealed bachelors’ institutions had the significantly highest ratio, \( \text{Mdn} = 0.68, \ M = 1.43, \ SD = 0.78, \text{Mdn} = 0.68 \) and compare with master’s \( M = 1.0, \ SD = 0.87, \text{Mdn} = 0.94 \), which also were significantly different from each other. In other words, the number of research opportunities did not grow proportionate to the number of faculty. The ratio of research courses per faculty also was significantly higher in private master’s and doctoral institutions \( M = 1.0, \ SD = 0.74 \) as compared with public schools \( M = 0.53, \ SD = 0.64 \), \( t(101) = 3.4, \ p < .001, \ r = .32 \).

**Required Research for Undergraduates, Content Area, and Institutional Differences**

Of the 203 responding departments, 98% \( (n = 199) \) offered at least one course with research opportunities and 160 required such a course \( (79\%) \). We emphasize that 21% of departments required no research of their majors. These findings support Cooney and Griffith’s (1994) finding that 72% of psychology departments required a research experience. Removing 2-year schools, 77% \( (n = 139) \) of the 181 remaining departments required a course with research. For the 19 \( (of 22) \) 2-year schools with courses with research opportunities, 63% \( (n = 12) \) required at least one research experience; overall, 55% did so. For bachelor’s schools, 43 of 51 that offered at least one course required at least one \( (84\%) \), overall, 73% did so. All responding master’s and doctoral departments offered at least one course with research. For master’s departments, 86% \( (50 \ of 58) \) with such a course required at least one, for doctoral 73% \( (47 \ of 64) \).

Overall, 39% \( (n = 79) \) of all departments required a single course with research and 40% \( (n = 81) \) more than one. Twenty-seven percent \( (6 \ of 22) \) of associate schools required more than one, as did 36% \( (21 \ of 59) \) of bachelor’s, 64% \( (37 \ of 58) \) of master’s, and 27% \( (17 \ of 64) \) of doctoral. Kulik (1973) found that 65% of departments required more than one laboratory course, substantially more than the 40% of departments in our study requiring two or more courses with any type of research experience. Perhaps expansion and changes of courses most frequently listed by departments (Perlman & McCann, 1999a) are partially responsible for the difference in estimates.

Table 2 presents data on required courses with research, as compared with all research courses, and their content areas by institutional type. Overall, 27% \( (n = 328) \) of all courses with research were required for the major—64% \( (n = 216) \) were in the methodology/experimental area.

**Types of Research Opportunities**

Table 3 presents data on five types of research opportunities by institutional type. Overall, 88% of departments offered a class with structured research and 90% classes with research in nonlaboratory settings. In 95% of the departments students could do individual research projects that they designed and conducted, whereas working collaboratively with faculty (65%) or graduate students (27%) was available less often. Public master’s and doctoral universities offered significantly more of these opportunities \( M = 4.1, SD = 0.94 \) than private schools \( M = 3.6, SD = 1.02 \), \( t(120) = 2.6, \ p < .009 \) \( (two-tailed) \), \( r = .23 \). Departments in bachelor’s, master’s, and doctoral institutions were comparable on structured and nonlaboratory research opportunities but the percentage requiring research experiences varied widely.

**The First Course with Research Opportunities**

**Content area**. For almost two thirds of responding departments, courses in the methodology/experimental area provided the first research opportunity. There were almost no institutional differences for the first research course. For an-

---

### Table 2. Required Courses With Research Opportunities for Majors in Seven Content Areas

<table>
<thead>
<tr>
<th>Content Area</th>
<th>Institutional Type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Associate</td>
</tr>
<tr>
<td>Introductory</td>
<td>89.8</td>
</tr>
<tr>
<td>Methodology/experimental</td>
<td>50.5</td>
</tr>
<tr>
<td>Social/development</td>
<td>42.5</td>
</tr>
<tr>
<td>Natural science</td>
<td>50.0</td>
</tr>
<tr>
<td>Clinical</td>
<td>0.0</td>
</tr>
<tr>
<td>Applied/vocational</td>
<td>0.0</td>
</tr>
<tr>
<td>Advanced/special topics</td>
<td>33.1</td>
</tr>
<tr>
<td>Total</td>
<td>50.0</td>
</tr>
</tbody>
</table>

\*Number of courses with research that are required divided by all courses in the content area offered with research (elective or required). For example: statistics, experimental, research methods, tests and measurement. For example: social, child, developmental, human growth/development, adolescent, adult development, exceptional child. For example: industrial/organizational, counseling, field experience, educational. For example: adjustment, abnormal, personality. For example: learning, cognitive, physiological, biological, sensation and perception, motivation. For example: independent study, special topics, seminar in psychology, research participation.
other 17%, introductory psychology provided this experience. In doctoral institutions, majors often waited for advanced and special topic courses to practice psychological science (13% of the first courses).

Course level. How soon in their education can students do research? Over all institutional types, 25% of the departments had first-research-opportunity courses during the first year, 47% the second, 23% the third, and 4% during the senior year. In over one-fourth of the departments, students’ undergraduate education was half completed before they had the possibility of doing psychological science. All departments for which introductory psychology was a student’s first such opportunity offered the course during the freshman and sophomore years. Most departments (70%) where the methodology/experimental area offered the first research opportunity had courses at sophomore and junior levels. Students must wait until their junior or senior year for a course with research experiences in 19% of bachelor’s departments, 44% of master’s, and 27% of doctoral.

Undergraduate Thesis/Major Project

If a department required a thesis or major project, students usually did research. We asked respondents if their department required an undergraduate thesis or substantive project (senior experience) for majors, telling them to answer “no” if the required senior capstone seminar had no “large” project. Of the 181 questionnaires returned for non–2-year schools, 172 responded. Overall, 26% (n = 45) of departments had such a requirement. Thirty-eight percent of 4-year schools offered such a course (n = 22), 34% of master’s (n = 19), and 7% of doctoral institutions (n = 4). Of these 45 departments, 40 provided data on what percentage of majors included data collection and analysis. For 4-year schools, 80% (SD = 30.4) of required senior theses or major projects included research with data collection and analysis, 82% (SD = 30.5) for master’s institutions, and 100% for the two doctoral universities.

Characteristics of Courses with Research Experiences

We divided the data on specific courses into two categories. Because so few courses in the clinical content area (n = 40) offered research opportunities, we excluded them. We included the few introductory courses with such experiences because of interest in introductory research opportunities (e.g., Berthold et al., 2003). Table 4 displays information on “curricular dimensions,” and Table 5 summarizes the type of research students did.

Departments most often required research experiences in methodology/experimental courses, but 34% of these courses were not required for the major. The psychology minor usually required no course with research opportunities, a qualitative loss in our opinion. The junior and senior advanced/special topics area offered numerous research opportunities, but faculty received teaching credit in only 49% of these courses. Many involved one-on-one or very small groups taught “above load.”

A better indicator of availability of research opportunities than total number of courses is the number of course sections and their enrollment, in other words, comparing number of seats with number of majors. We concluded there were sufficient seats in courses with research experiences when comparing the median data on numbers of majors and sections and median enrollment for the three types of institutions. However, research is secondary to content in most of these courses (the natural science and advanced/special topics areas notwithstanding). Only in the advanced/special topics area was research the primary emphasis most of the time (89%).

The few introductory courses with research emphasized course content, with students’ research almost always a secondary goal. Research was the primary goal in only 31% to 35% of the methodological/experimental courses for non–2-year departments. Departments seldom required majors to take courses in the natural science area with research, and research was the primary goal for those courses mostly for departments in doctoral institutions (60%) as compared with 34% of bachelor’s and 41% of master’s. Applied/vocational courses were junior and senior level, seldom required for the major, and faculty did not always receive credit for teaching them. Advanced/special topics courses were the courses for which faculty were least likely to receive teaching credit (45% of courses at bachelor’s, 30% at master’s, 65% at doctoral were taught “above load”). Paradoxically, these courses were most likely to have research as their primary goals (88% or higher at all de-

---

Table 3. Five Types of Research Opportunities Available for Undergraduate Psychology Majors

<table>
<thead>
<tr>
<th>Type of Research Opportunity</th>
<th>Institutional Type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Associate</td>
</tr>
<tr>
<td></td>
<td>%</td>
</tr>
<tr>
<td>Classes with structured research (e.g., labs)</td>
<td>27</td>
</tr>
<tr>
<td>Classes with research in nonlab settings (e.g., field, observational)</td>
<td>68</td>
</tr>
<tr>
<td>Individual research projects designed/conducted by students</td>
<td>22</td>
</tr>
<tr>
<td>Assist or work collaboratively with faculty</td>
<td>5</td>
</tr>
<tr>
<td>Assist or work collaboratively with grad student(s)*</td>
<td>—</td>
</tr>
</tbody>
</table>

Note. Associate: n = 22, bachelor: n = 59, master: n = 58, doctoral: n = 64. E = elective course; R = required for major.

*2-year schools that have no graduate students were omitted.
apartments), and median enrollments were low (10 at bachelor’s and master’s, 20 at doctoral).

**Types of Research Experiences Offered**

Table 5 presents types of research experiences by content area. In-class exercises based on a teacher’s ideas were common (35% of courses), and in 30% the instructor generated ideas and design for more extensive projects. Computer simulated (“virtual”) research was most common in the introductory, methodology/experimental, and natural science areas but uncommon elsewhere. Respondents indicated that computer simulation was seldom used by itself to teach psychological science. Instructors made most use of pre-existing data in the methodology/experimental area.

In only 45% of the courses were students fully responsible for all steps in their research, most often in the methodology/experimental and advanced/special topics areas. On average, 40% of all courses had a research project of less than 2 weeks’ duration, suggesting that students experience a lot of very brief research. Students had opportunities for collaborative or interdisciplinary research with biology and other disciplines in 11% of the courses, but we do not know the percentage of courses in which or students for whom this opportunity was realized. In the natural science area, bachelor’s settings were most likely to have collaborative research possibilities (13% of courses). Research experiences in groups were quite prevalent, and most data came from human participants. Animal use was most prevalent in natural science (40%) and advanced/special topics courses (29%).

Finally, we were interested as to whether and how students shared their science (made it public), a requisite for true science (American Psychological Association, 2001). Turning in a paper was the most popular avenue (64% of courses), 23% of courses used oral in-class presentations, and 11% used poster sessions. In 1 of 10 courses instructors used “other” ways (e.g., oral defense, college symposium). Respondents’ comments indicated that some faculty worked with students on presentation of their work at student research conferences and appropriate journal submission and publication, but these activities were seldom required and usually took place after the course had ended. We have no accurate count of such activities.

**Future Research Opportunities**

We asked respondents if, over the next 5 years, the number of research opportunities in psychology courses at their institution would “decrease,” “stay the same,” or “increase.” Sixty percent of faculty believed research opportunities would increase. There were no substantive differences between types of institution.
Discussion

Improving Students’ Psychological Science Experiences

McKeachie and Milholland’s (1961) concerns about the tension between adequate content coverage and an emphasis on research, coordinating laboratory work and content, and the range and complexity of experiments remain cogent. If doing science is as important in learning psychology and the scientific method as has been argued (e.g., Brewer et al., 1993; Committee on Undergraduate Science Education, 1997; Elmes, 2002, Halonen et al., 2003) and if psychology faculty are serious about preparing students for the world that awaits them, we can suggest desirable changes. All departments must require research of their majors (presently 21% do not do so). Departments need to increase the number of students fully responsible for their research at some point in their education, the number of courses in which research is the primary goal, the number of such courses required for the major, and the number of opportunities for longer research experiences. Research opportunities should be available earlier than the junior year.

To serve all of their majors and introduce students to a variety of scientific methods, departments must balance experimental research and psychometric inquiry and guard against concentrating too many research courses and experiences in the advanced/special topics area. Nonetheless, Brewer et al. (1993) listed students doing research with faculty collaboratively from conceptualization to presentation or publication as a defensible capstone experience, but few departments have chosen this option.

We cannot ignore the issue of the zeitgeist in higher education. There are many rewards for faculty and teaching competes with scholarship in its many forms and with service. Teaching of psychological science may be even more inten-

Table 5. Psychology Undergraduate Research Experiences

<table>
<thead>
<tr>
<th>Research Experience</th>
<th>Introductory</th>
<th>Methodology/Experimental&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Social/Developmental&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Natural Science&lt;sup&gt;c&lt;/sup&gt;</th>
<th>Applied/Vocational&lt;sup&gt;d&lt;/sup&gt;</th>
<th>Advanced/Special Topics&lt;sup&gt;e&lt;/sup&gt;</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instructor generates ideas for in-class exercise that results in data</td>
<td>60</td>
<td>46</td>
<td>43</td>
<td>59</td>
<td>11</td>
<td>8</td>
<td>35</td>
</tr>
<tr>
<td>Instructor generates idea and design for larger projects</td>
<td>23</td>
<td>26</td>
<td>23</td>
<td>30</td>
<td>8</td>
<td>23</td>
<td>30</td>
</tr>
<tr>
<td>Computer simulation</td>
<td>26</td>
<td>29</td>
<td>7</td>
<td>25</td>
<td>4</td>
<td>5</td>
<td>16</td>
</tr>
<tr>
<td>Students analyze pre-existing data “Idea bank” used, students refine data and fully responsible for research</td>
<td>21</td>
<td>35</td>
<td>19</td>
<td>17</td>
<td>7</td>
<td>13</td>
<td>20</td>
</tr>
<tr>
<td>Students responsible for all steps in research</td>
<td>6</td>
<td>10</td>
<td>10</td>
<td>13</td>
<td>2</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>At least one research project extends beyond 2 weeks’ duration</td>
<td>23</td>
<td>47</td>
<td>29</td>
<td>32</td>
<td>19</td>
<td>69</td>
<td>45</td>
</tr>
<tr>
<td>Data gathered</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>On campus (e.g., lab)</td>
<td>47</td>
<td>68</td>
<td>47</td>
<td>83</td>
<td>12</td>
<td>88</td>
<td>69</td>
</tr>
<tr>
<td>Off campus (e.g., field research)</td>
<td>13</td>
<td>29</td>
<td>50</td>
<td>14</td>
<td>82</td>
<td>45</td>
<td>37</td>
</tr>
<tr>
<td>Collaborative/interdisciplinary research (e.g., with biology)</td>
<td>15</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>20</td>
<td>11</td>
</tr>
<tr>
<td>Students work individually</td>
<td>64</td>
<td>66</td>
<td>48</td>
<td>55</td>
<td>30</td>
<td>85</td>
<td>64</td>
</tr>
<tr>
<td>In groups</td>
<td>45</td>
<td>55</td>
<td>43</td>
<td>48</td>
<td>19</td>
<td>19</td>
<td>38</td>
</tr>
<tr>
<td>Source of data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self</td>
<td>40</td>
<td>24</td>
<td>12</td>
<td>24</td>
<td>5</td>
<td>6</td>
<td>16</td>
</tr>
<tr>
<td>Classmates</td>
<td>45</td>
<td>49</td>
<td>18</td>
<td>33</td>
<td>5</td>
<td>21</td>
<td>30</td>
</tr>
<tr>
<td>Other human participants</td>
<td>30</td>
<td>61</td>
<td>88</td>
<td>54</td>
<td>86</td>
<td>59</td>
<td>63</td>
</tr>
<tr>
<td>Animals</td>
<td>4</td>
<td>12</td>
<td>1</td>
<td>40</td>
<td>3</td>
<td>29</td>
<td>20</td>
</tr>
<tr>
<td>Data made public domain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher requires paper</td>
<td>63</td>
<td>65</td>
<td>58</td>
<td>68</td>
<td>48</td>
<td>67</td>
<td>64</td>
</tr>
<tr>
<td>Course poster session</td>
<td>2</td>
<td>17</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>15</td>
<td>11</td>
</tr>
<tr>
<td>Course oral presentation</td>
<td>19</td>
<td>26</td>
<td>24</td>
<td>21</td>
<td>13</td>
<td>26</td>
<td>23</td>
</tr>
<tr>
<td>Other&lt;sup&gt;f&lt;/sup&gt;</td>
<td>9</td>
<td>9</td>
<td>5</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>10</td>
</tr>
</tbody>
</table>

Note. Data are based on the number of courses in each content area. Introductory n = 47, experimental/methodology n = 370, social/developmental n = 151, applied/vocational n = 119, natural science n = 271, advanced/special topics n = 397. We omitted the Clinical area—there were only 37 courses.

<sup>a</sup>For example: statistics, experimental, research methods, tests and measurement. <sup>b</sup>For example: social, child, developmental, human growth/development, adolescent, adult development, exceptional child. <sup>c</sup>For example: industrial/organizational, counseling, field experience, educational. <sup>d</sup>For example: learning, cognitive, physiological, biological, sensation and perception, motivation. <sup>e</sup>For example: independent study, special topics, seminar in psychology, research participation. <sup>f</sup>Includes presentations at an oral or honors defense, college symposium or other forums, regional or state psychology conferences, publication, or placing the student research on an instructor’s or department’s Web site.
sive and consuming than most other forms of pedagogy, and the discipline may need to explore ways to develop and interest teachers in students doing science. Establishment of “science teaching” awards and other recognitions for faculty successful in this arena and workshops on this topic at regional and national meetings to increase visibility of students learning science and teachers’ expertise and interest may be desirable. The Society for the Teaching of Psychology could develop a subgroup of teachers dedicated to students actively learning scientific processes. The discipline, through the American Psychological Association (APA) and the American Psychological Society (APS), could advocate for and educate about such pedagogical scientific efforts.

As the Committee on Undergraduate Science Education (1997) noted:

It is hard to imagine learning to do science, or learning about science, without doing laboratory or fieldwork. Experimentation underlies all scientific knowledge and understanding. Laboratories are wonderful settings for teaching and learning science. They provide students with opportunities to think about, discuss, and solve real problems. (p. 16)

Uses for the Data Including Assessment and Policy Purposes

Our data can serve as a future baseline for changes in how and the quality with which the discipline teaches science and for comparing psychology to other sciences. Departments can compare their current curriculum with these data and recommendations, and decide if changes are needed. They may want to assess the outcome of such experiences either informally or integrate such assessment in their formal curricular assessments. The data also identify lacunae in the literature. For example, over 10% of all courses offered interdisciplinary research opportunities, and yet in reviewing the journal Teaching of Psychology from 1993 to the present, we found only two interdisciplinary courses described (Albrecht & Nelson, 2001; Misale, Gillette, & delMas, 1996). More accounts are needed to call attention to this type of teaching.

The APA’s Board of Educational Affairs’ learning goals and outcomes (Halonen et al., 2003) provides a context and structure for high quality student learning experiences emphasizing students learning psychological science. The complexity of the educational landscape in psychology, including variations in schools’ and departments’ selectivity, curricular goals, and the experience and commitment of its teachers, make assessment and consensus on meeting these goals difficult. Nonetheless, our data provide one avenue for discussion and for developing curricular and course exemplars for student practice of psychological science. They also could assist in the development of approval and certification guidelines for undergraduate education in psychology if the discipline chooses to emulate chemistry’s model. Additionally, the discipline should gather and integrate data on the relation of students’ pedagogical experiences and their preparation for future education and careers. In our department, a simple exit questionnaire for all seniors provides data on the curriculum’s most useful course(s) and what students found most valuable in being a major. It would be relatively easy to expand this effort.

Future Research

There are a variety of research projects that would increase the knowledge about students doing psychological science. First, we wish we had inquired about research being several weeks long or even longer to better assess the depth and nature of such experience. Second, do students’ attitudes toward psychological science become more positive, and their critical thinking, fluency with numbers, and other dimensions actually increase as an outcome of their research experiences? Our research did not address this critical question. Third, our experience as faculty tells us that a quality science education is not necessarily a consequence of the resources and number of people a department puts into this endeavor (although these factors contribute), but of the quality of teaching and nature of the research experiences. What are the critical dimensions related to meaningful student engagement with psychology’s scientific method?

Fourth, how will the discipline’s learning goals and outcomes (Halonen et al., 2003) influence availability and quality of research opportunities for undergraduates? Fifth, how prevalent is the practice of faculty working with students to submit, present, or publish their research after the formal course in which this research was conducted is completed, and what are the educational benefits for students?

Sixth, although our response rate exceeded 40%, only a few responses could create a 10% difference between departmental types. Replicating and expanding the research in the future will yield more reliable conclusions about institutional differences. That is one reason we hesitate to draw too many conclusions about differences between bachelor’s, master’s, and doctoral institutions. Each type of institution had relative strengths as compared to the other two in some facets of the practice of research for undergraduates. For example, the large faculties in departments in doctoral universities may not translate proportionately into more undergraduate research experiences. These departments often have large numbers of majors and other undergraduates enrolling in their courses as well as time-consuming graduate responsibilities. One doctoral respondent noted that there were so many students and so few faculty willing to teach the lab experience that multiple-lab courses were condensed into a generic lab only for the Bachelor of Science student. Similarly, most significant differences for public versus private institutions reflected the larger number of majors and faculty in public schools, not their funding identity.

Conclusions

An incongruency exists between what psychologists have proclaimed and continue to say about the importance of having undergraduates practice psychological research and typical curricula and courses. The nature of the undergraduate curricula, strong in so many ways, needs to incorporate more active learning of science for psychology students if the teachers of psychology wish to maximize the important and varied outcomes of these experiences. The American Chemical Society’s educational guidelines (Committee on Professional Training, 2003) set the standard and provide a useful...
model, and APA’s recent attention to undergraduate educational goals (Halonen et al., 2003) may provide momentum for both APA’s and the APS’s continued attention and advocacy for undergraduate learning about and practice of the science of psychology.

References


Notes

1. The research was supported by a grant from the Office of Teaching Resources in Psychology of the Society for the Teaching of Psychology.

2. This article is based on a poster presentation at the 2004 National Institute on the Teaching of Psychology, St. Petersburg Beach, FL.

3. We gratefully acknowledge Barney Beins, Charles L. Brewer, Bill Buskist, David Elmes, and Thomas McGovern for their assistance and comments. A special thanks to our two research assistants, Brian Ayotte and Brad Brzozowski. The research could not have been conducted without the commitment and assistance of the participants who completed questionnaires. We thank Dr. Beins and APA for providing mailing labels.

4. Send correspondence and requests for copies of the questionnaire and a full technical report to Baron Perlman, Department of Psychology, University of Wisconsin, Oshkosh, WI 54901; e-mail: Perlman@uwosh.edu.
APA's Learning Objectives for Research Methods and Statistics in Practice: A Multimethod Analysis
Thomas J. Tomcho, Diana Rice, Rob Foels, Leah Folmsbee, Jason Vladescu, Rachel Lissman, Ryan Matulewicz and Kara Bopp
*Teaching of Psychology* 2009 36: 84
DOI: 10.1080/00986280902739693

The online version of this article can be found at:
http://top.sagepub.com/content/36/2/84

Published by:

SAGE
http://www.sagepublications.com

On behalf of:

Society for the Teaching of Psychology

Additional services and information for *Teaching of Psychology* can be found at:

Email Alerts: http://top.sagepub.com/cgi/alerts

Subscriptions: http://top.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Apr 1, 2009

What is This?
APA’s Learning Objectives for Research Methods and Statistics in Practice: A Multimethod Analysis

Thomas J. Tomcho
Salisbury University

Diana Rice
Geneva College

Rob Foels
Amherst College

Leah Folmsbee, Jason Vladescu, and Rachel Lissman
Syracuse University

Ryan Matulewicz
York University

Kara Bopp
Wofford College

Research methods and statistics courses constitute a core undergraduate psychology requirement. We analyzed course syllabi and faculty self-reported coverage of both research methods and statistics course learning objectives to assess the concordance with APA’s learning objectives (American Psychological Association, 2007). We obtained a sample of 64 research methods and 56 statistics syllabi (return rates: 16%, 14%) from 400 randomly selected psychology departments in colleges and universities in the United States. Course syllabi generally contained appropriate research methods and statistics content. However, certain APA learning objectives consistently were absent from course syllabi, and we found discrepancies between listed objectives and faculty self-reports of topics covered.

Across the psychology curriculum, instructors include explicit learning objectives to clarify course goals for students. Explicit learning objectives can be particularly useful in traditionally challenging courses such as research methods (Perlman & McCann, 1999a, 1999b) and statistics (Bartz, 1981; Gordon & Gordon, 1992). In contrast to the many articles addressing teaching techniques for research methods (e.g., Boyce & Geller, 2002; Dunn, 2000; Isbell & Tyler, 2003; O’Dell & Hoyert, 2002; Sommer & Sommer, 2003) and statistics (e.g., Marasinghe, Meeker, Cook, & Shin, 1996; Perkins & Saris, 2001; Radke-Sharpe, 1991; Stroup, Goodman, Cordell, & Scheaffer, 2004; Warner & Meehan, 2001), researchers have not published much research concerning the learning objectives used in typical undergraduate courses in research methods or statistics. Generally, researchers have examined what instructors view as essential topics in research methods and statistics (e.g., Giesbrecht, Sell, Scialfa, Sandals, & Ehlers, 1997). However, empirical data that examine learning objectives as operationalized by research methods and statistics course instructors are lacking. Therefore, the first goal of our research was to identify what research methods and statistics learning objectives teachers deem important by examining learning objectives as listed on syllabi and as self-reported by course instructors.

We also were interested in the methodological and statistical learning objectives outlined by the American Psychological Association (APA, 2007) for undergraduate education and whether instructors are listing
consonant course learning objectives. The APA outlined six learning subgoals for research methods comprised of 18 specific methods and statistics learning objectives that undergraduates should meet on completion of their degrees. These objectives range from being able to identify and articulate differences among various research designs to addressing the sociocultural constraints on generalizability of research findings. Therefore, a second goal of our research was to identify the degree to which instructors’ learning objectives were concordant with the APA’s stated objectives for undergraduate research methods and statistics education.

Method

Sample Identification and Recruitment

We used Peterson’s Guides (2002) to 4-year colleges and universities to identify the population of 4-year colleges and universities in the United States that offered a psychology degree. We categorized the 1,228 American institutions offering undergraduate psychology majors by one of seven Carnegie types (Carnegie Foundation, 2000). We then selected a proportionate stratified random sample of 400 institutions based on the percentage of each Carnegie type institution to the overall total. We randomly selected the following numbers of schools by Carnegie type: I (12% of total, n = 48), II (7%, n = 28), III (37%, n = 148), IV (7%, n = 28), V (16%, n = 64), VI (20%, n = 80), and VII (1%, n = 4).

We mailed surveys to department chairs at identified institutions asking that they forward the research methods survey to the faculty member with primary responsibility for teaching the departmental course in research methods and the statistics survey to the faculty member with primary responsibility for teaching the departmental course in statistics. At those institutions that might have had multiple instructors teaching research methods and statistics, we relied on the department chairs’ judgment regarding who should complete the survey. As part of a larger study of undergraduate instruction in psychology, we asked research methods and statistics course instructors to return two items in a postage-paid return envelope: (a) a copy of their most recent syllabus for content analysis of course learning objectives, and (b) a brief teaching survey regarding APA’s 18 learning objectives for research methods and statistics in psychology. The survey asked faculty to rate the amount of time during the course devoted to each of the 18 APA learning objectives for research methods and statistics based on the following scale: 0 = never covered; 1 = covered less than 20% of course time; 2 = covered between 20% and 40% of course time; 3 = covered 40% to 60% of course time; 4 = covered 60% to 80% of course time; 5 = covered greater than 80% of course time.

After 6 weeks we sent follow up e-mail requests for participation to increase our response rate. In addition, we conducted a Web-based search for syllabi used by nonrespondents in our sample. With respect to course syllabi, follow-up e-mails and Web-based searches yielded a final sample of 64 research methods syllabi for an approximate 16% return rate; of these 64, 53% (n = 34) also completed the teaching survey. The same process yielded a final sample of 56 statistics syllabi for an approximate 14% return rate; of these 56, 39% (n = 22) also completed the teaching survey. Two other statistics instructors also completed the teaching survey but did not provide syllabi for analysis. Because of the relatively low return rate and the similarities in syllabi content across Carnegie types, we aggregated results across institutions for both research methods and statistics syllabi.

Content Analyses

We used APA’s 18 specific learning objectives (APA, 2007) for research methods as a conceptual framework for categorizing course learning objectives. Based on information provided by instructors on their syllabi, we also included the development of computer-based statistical skills as an additional 19th learning objective. Thus, we used 19 specific learning objectives to categorize course learning objectives.

We conducted several group meetings prior to, and during the process of, categorizing course learning objectives to train coders and protect against coder drift. Two independent advanced undergraduate coders parsed course learning objectives listed on syllabi into unitary objectives. For example, an objective listed as “To teach students how to use statistics to critically analyze the research of others and to incorporate appropriate techniques when designing their own research projects” was categorized as “Use appropriate statistical strategies to collect data, analyze data, interpret data, and report research” (APA Learning Objective 12). After we parsed syllabi course learning objectives into unitary components, two independent coders next categorized each unitary component into one of the 19 different learning objectives. During the
initial period, coders worked together coding syllabi to facilitate agreement on how to identify which category a specific course learning objective represented. Subsequently, during the independent coding of course learning objective unitary components, if questions about potential categorization arose (which was infrequent), a conference discussion among the lead authors clarified the specific relevant category for a particular course learning objective. Two coders rated a random sample of syllabi, resulting in acceptable levels of intercoder reliability ($r = .87$).

Results

We report our results in two sections. We first report on learning objectives for research methods and then for statistics. Within each section, we report the results of content analyses of course syllabi followed by self-report responses from the teaching questionnaire. We present in Table 1 the percentage of syllabi in which instructors listed each APA objective as a course objective. Further, we indicate in Table 1 the amount of time that faculty self-reported allotting to each objective. Please note that percentages for faculty self-report might exceed 100% in those instances where faculty addressed course objectives in multiple different contexts.

Learning Objectives: Research Methods

Research methods syllabi. Table 1 shows that the majority of instructors indicated at least four of the learning objectives on their syllabi, but few, if any, instructors indicated emphasizing objectives that involve understanding cautions and limits to applying statistical information broadly without considering individual differences and sociocultural contexts.

Research methods teaching survey. Table 1 shows that the majority of instructors indicated at least half (8 of 18) of APA’s learning objectives.

Learning Objectives: Statistics

Statistics syllabi. Table 1 shows that the majority of instructors indicated at least two of the learning objectives on their syllabi. However, no instructors indicated on their syllabi learning objectives that address the kinds of cautions and limits to broad application of statistical information that the APA (2007) considered important.

Statistics teaching survey. Table 1 shows a general pattern of agreement existing between instructors’ report of time devoted to learning objectives and the learning objectives listed on syllabi. However, there was one discrepancy in this general pattern concerning cautions and limitations to the broad application of statistical results. That is, instructors reported 20% to 40% of class time devoted to the coverage of these objectives, yet they received no mention on syllabi.

Discussion

We examined undergraduate research methods courses and statistics courses with two goals in mind: (a) to report on instructors’ learning objectives in these courses, and (b) to report on the degree to which instructors’ learning objectives were concordant with APA’s learning objectives for research methods skills. We used a multimethod approach to accomplish these goals by content analyzing course syllabi and eliciting self-report data from course instructors.

Regarding learning objectives for research methods, we found that the appropriate use of various research designs and statistical strategies and the use of relevant research, theory, and APA ethics in the planning and interpretation of research were prominent in both syllabi and in instructor self-reports. Regarding learning objectives for statistics, we found that the interpretation of basic statistical results, the use of appropriate statistical strategies, the evaluation of the validity of conclusions presented in research, and the development of proficiency with statistical software packages were prominent in both syllabi and instructor self-report. We found that for both research methods and statistics syllabi, instructors listed learning objectives that were a subset of APA’s learning objectives. However, important qualifications to the basic findings emerged that warrant further attention: (a) Course syllabi for research methods and statistics consistently lacked key APA learning objectives, and (b) discrepancies existed between course syllabi and instructor self-report.

From our examination of both research methods and statistics course syllabi, there were nine APA learning objectives that instructors either failed to mention or that received only minimal coverage based on objective data obtained from these syllabi. Fewer than 10%
Table 1. Course Learning Objectives

<table>
<thead>
<tr>
<th>Objectives</th>
<th>Indicated on Syllabi</th>
<th>Faculty Self-Report</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Research Methods %</td>
<td>Statistics %</td>
</tr>
<tr>
<td>1. Describe the basic characteristics of the science of psychology</td>
<td>53%</td>
<td>16%</td>
</tr>
<tr>
<td>2. Describe how various research designs address different types of</td>
<td>70%</td>
<td>5%</td>
</tr>
<tr>
<td>questions and hypotheses</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Articulate strengths and limitations of various research designs</td>
<td>20%</td>
<td>5%</td>
</tr>
<tr>
<td>4. Distinguish the nature of designs that permit causal inferences from</td>
<td>9%</td>
<td>5%</td>
</tr>
<tr>
<td>those that do not</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Interpret basic statistical results</td>
<td>27%</td>
<td>80%</td>
</tr>
<tr>
<td>6. Distinguish between statistical significance and practical significance</td>
<td>5%</td>
<td>4%</td>
</tr>
<tr>
<td>7. Describe effect size and confidence intervals</td>
<td>0%</td>
<td>4%</td>
</tr>
<tr>
<td>8. Evaluate the validity of conclusions presented in research reports</td>
<td>20%</td>
<td>29%</td>
</tr>
<tr>
<td>9. Locate and use relevant databases, research theory to plan research,</td>
<td>61%</td>
<td>16%</td>
</tr>
<tr>
<td>conduct research, and interpret results of research reports</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Formulate testable research hypotheses, based on operational</td>
<td>3%</td>
<td>7%</td>
</tr>
<tr>
<td>definitions of variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. Select and apply appropriate methods to maximize internal and external</td>
<td>3%</td>
<td>18%</td>
</tr>
<tr>
<td>validity, and reduce plausibility of alternate explanations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. Use appropriate statistical strategies to collect data, analyze data,</td>
<td>53%</td>
<td>54%</td>
</tr>
<tr>
<td>interpret data, and report research</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. Recognize that theoretical and sociocultural contexts as well as</td>
<td>11%</td>
<td>2%</td>
</tr>
<tr>
<td>personal biases may shape research questions, research designs, data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>collection, data analysis, and data interpretation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14. Follow APA Code of Ethics in the treatment of human and nonhuman</td>
<td>66%</td>
<td>2%</td>
</tr>
<tr>
<td>participants in designing research, data collection, data interpretation,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>and reporting of psychological research</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15. Exercise caution in predicting behavior based on the limitations of</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>single studies</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16. Recognize the limitations of applying normative conclusions to</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>individuals</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17. Acknowledge that research results may have unanticipated societal</td>
<td>2%</td>
<td>0%</td>
</tr>
<tr>
<td>consequences</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18. Recognize that individual differences and sociocultural contexts may</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>influence the applicability of research findings</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19. Become proficient in the use of a statistical software package</td>
<td>22%</td>
<td>32%</td>
</tr>
</tbody>
</table>

Note. Faculty self-report: 0 = never covered, 1 = covered less than 20% of course time, 2 = covered between 20% and 40% of course time, 3 = covered 40% to 60% of course time, 4 = covered 60% to 80% of course time, 5 = covered greater than 80% of course time. Total percentages can exceed 100% in those instances where instructors might have addressed multiple course objectives in the context of the same course time period. Proficiency in statistical software packages was listed by instructors on their syllabi but was not asked about on the teaching survey. Software proficiency is not an APA learning objective for research methods, but is an APA learning objective for information and technological literacy.
of instructors mentioned Learning Objectives 4, 6, 7, 10, 13, 15, 16, 17, and 18 on course syllabi. Moreover, Learning Objectives 13 and 17, in addition to having little mention on syllabi, were not a focus of instructors’ class time, according to their self-reports. Thus, research methods and statistics teachers appear to be providing limited instruction pertaining to the sociocultural contexts or consequence of research and to cautioning students about the dangers of predicting behavior from single studies or applying normative findings to individuals.

In contrast to what course instructors expressly stated as learning objectives on their syllabi, based on self-reports they were more likely to cover certain topics more frequently. For example, only 20% of research methods syllabi included the expressly stated goal of articulating strengths and limitations of various research designs. In contrast, instructors self-reported nearly 80% of course time devoted to this topic. More dramatically, although only 9% of research methods syllabi included the expressly stated goal of distinguishing the nature of designs that permit causal inferences from those that do not, instructors on average self-reported 80% of course time devoted to this topic. Indeed, as shown in Table 1, there were great discrepancies between syllabi and survey results for several of the learning objectives. We noted similar findings for statistics. For example, no statistics syllabi included the expressly stated goal of exercising caution in predicting behavior based on the limitations of single studies, but instructors noted that they devoted 20% to 40% of class time to this topic.

What accounts for the discrepancies between what instructors list on their syllabi and what they indicate by self-report is unclear. Instructors might have used the teaching survey to indicate perceived importance of learning objectives rather than realistic allotment of course time. Thus they might have rated learning objectives highly on the survey without having them as explicit learning objectives on their syllabi. However, it also could be that instructors do not adequately articulate learning objectives on syllabi and, as such, provide instruction in more content areas than they list on their syllabi. If it is this latter case, educational theorists writing in the area of instructional theory (e.g., Gagné & Medsker, 1996) have suggested that instructors inform students of the learning objectives to frame students’ learning expectations (e.g., Driscoll, 2005).

APA recommends students obtain competencies in 18 research methods and learning objectives. However, our data indicate that research methods and statistics teachers limit their instructional coverage of such topics as the generalizability, limitations, and sociocultural contexts of research. It is plausible that teachers might believe either that there is already too much basic information to cover in an otherwise difficult research methods and statistics sequence or that they are addressing these learning objectives in other courses such as social psychology. Regardless, in recent years, APA and its members have placed increasing importance on the issues of sociocultural contexts and research consequences in research and teaching, and as such, instructors should provide more coverage of these objectives in research methods and statistics courses.

References


Notes

1. Thomas Tomcho collected these data while a faculty member at Syracuse University, Syracuse, NY.

2. Send correspondence to Thomas J. Tomcho, Department of Psychology, Salisbury University, Salisbury, MD 21802; e-mail: tjtomcho@salisbury.edu.
Using Empirical Article Analysis to Assess Research Methods Courses

Peter Bachiochi, Wendi Everton, Melanie Evans, Madeleine Fugere, Carlos Escoto, Margaret Letterman, and Jennifer Leszczynski

Abstract
Developing students who can apply their knowledge of empirical research is a key outcome of the undergraduate psychology major. This learning outcome was assessed in two research methods courses by having students read and analyze a condensed empirical journal article. At the start and end of the semester, students in multiple sections of an introductory research methods course and students in sections of an advanced methods course answered questions about a condensed journal article in a pre–post approach. Students in the advanced course significantly outperformed students in the introductory course at both administrations, and students in both courses improved significantly from beginning to end of the semester. Results indicate that using journal article analysis can effectively supplement assessment efforts for psychology departments.

Keywords
assessment, research methods, article analysis

Interest in the outcomes of undergraduate psychology education dates back to the 1970s (e.g., Lunneborg, 1974). Early approaches focused on gathering information about the outcomes of our teaching in order to satisfy the demands of key stakeholders (e.g., accrediting bodies, parents). Although these external demands remain, a continued and growing focus on the scholarship of teaching has converged with these external demands. As a result, a scientist–educator model, similar to the scientist–practitioner model, has emerged that “treats professional work as an inquiry into the effectiveness of practice” (Bernstein et al., 2010, p. 30). Rather than viewing assessment simply as a post hoc analysis, newer approaches focus on constant feedback and improvement. Ultimately, assessment efforts can satisfy stakeholders as well as contribute to a more scholarly approach to teaching and learning.

As a result, psychology programs have been studying their own effectiveness using a variety of methods to assess multiple student learning outcomes. To assess the breadth of learning outcomes thoroughly, multimethod approaches are required (Sheehan, 1994). Standardized tests (Stoloff & Feeney, 2002), capstone courses and senior seminars (Morgan & Johnson, 1997), and embedded assignments (Wade, 1995) are typical tools. Similarly, several different learning outcomes have been assessed. Critical thinking has been a consistent focus of existing research (Lawson, 1999; Wade, 1995), as has general knowledge of the discipline, often assessed by standardized tests (Stoloff & Feeney, 2002). Other researchers have addressed statistical knowledge and writing skills (Smith, 1995) as well as PsycLIT competence (Cameron & Hart, 1992), and Friedrich (1996) tested a measure of students’ perceptions of psychology as a science. Although far from an exhaustive list, the above research provides a glimpse into the wide variety of outcomes assessment activity in psychology programs.

The American Psychological Association (APA) took a preliminary step in identifying the major learning goals and objectives for undergraduate psychology majors (APA, Task Force on Undergraduate Psychology Major Competencies, 2002) that led to the APA Guidelines for the Undergraduate Psychology Major (APA, 2007). The guidelines outline 10 broad goals each with several learning outcomes. The first five goals/outcomes address “knowledge, skills, and values consistent with the science and application of psychology” and the second five address “knowledge, skills, and values consistent with liberal arts education that are further developed in psychology” (APA, 2007, p. 9). These guidelines provide a detailed road map for curriculum design and outcomes assessment for undergraduate psychology programs.

1 Eastern Connecticut State University, Willimantic, USA

Corresponding Author:
Peter Bachiochi, Department of Psychology, Eastern Connecticut State University, 83 Windham Street, Willimantic, CT 06226, USA
Email: bachiochip@easternct.edu
The second goal for the undergraduate psychology major outlined by APA is "Research Methods in Psychology." This goal arguably anchors the field of psychology as a science (Dunn et al., 2010). To achieve this goal at our institution we require a Psychology Statistics course and a two-course sequence in Research Methods: Research Methods I (RMI) and Research Methods II (RMII). Statistics and RMI are prerequisites for RMII, and although students are advised to take the Statistics course first, it is not a requirement. The RMI course addresses basic research designs (e.g., observational, correlational, experimental), ethical issues, APA format, and reading empirical research articles, and it provides a brief statistics overview/review. In this course, students also develop an independent research proposal as the culminating course project. In RMII, students spend more time engaging in critical analyses of research, gaining experience with a variety of research techniques, and fully completing their independent research project. Students conduct their proposed study by gathering and analyzing data, presenting the results (in poster or presentation format), and writing up the results in a complete APA-style empirical paper.

APA Goal 2 regarding research methods contains multiple subgoals, such as describing the basic characteristics of the science of psychology (2.1), explaining different methods (2.2), evaluating the appropriateness of conclusions derived from psychological research (2.3), designing and conducting basic studies (2.4), following the APA Code of Ethics (2.5), and generalizing research conclusions appropriately (2.6). We use the students’ independent research project in RMII to assess achievement of Subgoal 2.4. To assess whether the other subgoals are achieved through our research methods courses, we adapted the analysis of primary sources that Christopher and Walter (2006) described. Rather than using article analysis as a class exercise, we used an article analysis as a programmatic assessment tool. The intent of this test was to assess the students’ ability to identify key variables and design weaknesses (consistent with Subgoals 2.1 and 2.2), describe generalizability concerns (Subgoal 2.3), and discuss ethical issues (Subgoal 2.5). Questions also addressed students’ ability to critically analyze the conclusions of the authors (Subgoals 2.3 and 2.6).

Combining cross-sectional and longitudinal techniques, we took a value-added approach (Halpern, 1988) to assess our students’ skills as consumers of empirical research. We administered the test at the beginning and end of the semester in all sections of both courses to compare RMI students to RMII students (cross-sectionally) and to compare student performance from the beginning to the end of the respective courses (longitudinally):

**Hypothesis 1**: We predicted that at the start and end of the semester, the scores of students in RMII would be higher than the scores of those students taking RMI.

**Hypothesis 2**: We predicted that the scores of students taking RMI would increase significantly from the start of the semester (Time 1) to the end of the semester (Time 2).

We also predicted that the scores of students in RMII would increase over time.

**Hypothesis 3**: We predicted that students who had completed more psychology classes prior to this assessment would have higher scores, both in Time 1 and Time 2.

**Method**

**Participants**

Participants were students enrolled in either research methods course offered by the Psychology Department during one term at Eastern Connecticut State University. There were five sections of RMI and three sections of RMII. The instrument was administered during class time at the beginning and again at the end of the semester. In the first administration of the instrument, data were collected from 160 students (114 in RMI, 46 in RMII). Of those reporting demographic information, 89.8% were women, 88.5% Caucasian, and 5.1% were African American. These demographic characteristics generally match the demographics of the department’s majors. For the second administration, 130 students (88 in RMI, 42 in RMII) completed the materials. We do not have second administration data for 30 students because of a combination of some students not attending class during the administration and some students withdrawing from their course. To match the data for each participant across the two time periods, we used unique ID numbers, which yielded 128 matched pretest–posttest score pairs (87 score pairs for RMI and 41 score pairs for RMII).

**Measures**

The instrument was a primary research article by Powell and Drucker (as condensed in Lomand, 2002) and a set of questions we developed about the article (including two questions from the Lomand text). The Lomand (2002) text consists of a variety of articles that have been abbreviated for use as a class supplement. The articles contain all the sections of a typical journal article, but the abbreviated versions retain the original content/ intent of the article and take less time to read. An abbreviated article was essential because administration needed to be completed in one class period. The Powell and Drucker article was selected due to the potential interest of the topic to students (riding with a drunk driver), statistical sophistication consistent with student abilities (primarily chi-square tests), and its brevity, all consistent with the recommendations of Carkenord (1994) and others. It was also an article that had not been used by instructors in other courses.

Members of the department developed questions to assess some departmental learning goals that are consistent with the APA guidelines. There were 11 questions, designed to assess the students’ ability to identify key variables, ethical issues, generalizability concerns, and design weaknesses, as well as questions that asked students to critically analyze the conclusions of the authors. The questions also varied in level of difficulty to avoid floor and ceiling effects. During the first administration, we also collected demographic information,
such as which psychology courses students had completed, whether they had transferred to our school, full-time versus part-time status, and whether English was their primary language.

We developed a scoring rubric to increase the reliability and validity of our ratings. Some questions were worth a maximum of 1 point, while others were worth a maximum of 4 points. For example, 1-point questions addressed factual information such as “At what probability level was the relationship between the IV and the DV statistically significant?” and 4-point questions required more involved analysis, such as “Describe (don’t simply name them) two methodological weaknesses of the design other than generalizability.” Students could earn partial credit on questions (e.g., 3 points on a 4-point question). Overall, the range of possible scores on the instrument was 0 to 29 points. The 11 questions, APA guideline linkages, and question weights are provided in Table 1.

Procedure

The pretest information was collected during the first week of the academic term, and the posttest information was collected during the 13th week of a 15-week term. The class instructor collected the data during a regular class period. Students needed between 20 and 30 minutes to complete the instrument.

After both administrations were completed, each student’s answers were rated individually by two faculty members. Seven pairs of raters provided scores. To help reduce rater bias, all raters received and graded instruments from students enrolled in both RMI and RMII and from both administrations. Raters were blind to both the class of the student and the time that the assessment was administered. Raters assigned a point value for every answer on the instrument and summed the values for a total score. Mean scores were then computed using the two raters’ total scores.

Ratings still contain some level of subjectivity; therefore, it was important to assess interrater reliability. We used intraclass correlation (ICC), which computes the variability in ratings that can be accounted for by participants versus raters (Shrout & Fleiss, 1979). As the ICC approaches 1, variability in scores is attributed more to participants than raters (i.e., higher interrater reliability). These correlations were excellent for both administrations using the cutoffs determined by Fleiss, Levin, and Paik (2003), where ICC values of .75 or higher demonstrate excellent agreement. The ICC correlations we obtained ranged from .75 to .92, which indicated excellent agreement between raters.

Results

To test our hypotheses we used a series of \( t \) tests with a Bonferroni adjustment to the \( p \) value. We chose not to use a \( 2 \times 2 \) analysis because this would have excluded the first administration data of the 30 students who were not present at the second administration due to the attrition mentioned above. First, we compared the scores on the article analysis of students taking RMII with students taking RMI using the independent samples \( t \) test. For the first administration, students in RMII had significantly higher scores \((M = 15.27, SD = 4.43)\) compared to students in RMI \((M = 11.47, SD = 4.47)\), \(t(158) = 4.88, p < .001, d = .776\). For Time 2, students in RMII had significantly higher scores \((M = 17.63, SD = 4.18)\) compared to students in RMI \((M = 14.75, SD = 4.44)\), \(t(128) = 3.52, p < .001, d = .622\).

For the second hypothesis, we matched the two test scores of each student to determine if scores increased from the beginning to the end of the semester. Note that the means are slightly different from those reported in the previous analyses because the repeated-measures test includes only the students who took the test twice. Using a repeated-measures \( t \) test we analyzed the results for RMI first and found that indeed scores had improved significantly from students’ scores in January \((M = 11.43, SD = 4.48)\) compared to the same students’ scores in April \((M = 14.81, SD = 4.42)\), \(t(87) = 7.53, p < .001, d = 1.14\). Scores also improved significantly for RMII students from January \((M = 15.32, SD = 4.64)\) to April \((M = 17.77, SD = 4.12)\), \(t(41) = 3.74, p < .001, d = .83\). A Bonferroni correction

<table>
<thead>
<tr>
<th>Question</th>
<th>APA Guideline Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Did the driver have a fake beer in his hand in all four conditions?</td>
<td>N/A</td>
</tr>
<tr>
<td>Identify the levels of the independent variable in this experiment.</td>
<td>2.1</td>
</tr>
<tr>
<td>How many of the 10 participants refused to enter the car in the condition where the confederate refused to enter the car?</td>
<td>2.3</td>
</tr>
<tr>
<td>At what probability level was the relationship between the IV and the DV significant?</td>
<td>2.3</td>
</tr>
<tr>
<td>Describe the way(s) deception was used in this study.</td>
<td>2.5</td>
</tr>
<tr>
<td>The researchers recommend gender as an interesting variable for future study. How might gender influence the results?</td>
<td>2.4, 2.6</td>
</tr>
<tr>
<td>Identify three factors that impact the generalizability of the results.</td>
<td>2.6</td>
</tr>
<tr>
<td>Describe (don’t simply name them) two methodological weaknesses of the design other than generalizability.</td>
<td>2.2</td>
</tr>
<tr>
<td>Can the reader conclude that peer conformity causes one to be more likely to ride with an intoxicated driver? Explain.</td>
<td>2.2, 2.3</td>
</tr>
<tr>
<td>Discuss the implications of this study’s results.</td>
<td>2.6</td>
</tr>
<tr>
<td>Write how this article would appear in an APA formatted reference page.</td>
<td>7.1^a</td>
</tr>
</tbody>
</table>

\( DV = \) dependent variable; \( IV = \) independent variable.

\(^a\) Part of Communication Skills goal and not Research Methods in Psychology.
(\( p = .025 \)) was used with all \( t \) tests to control for the risk of a Type I error.

Our final hypothesis was also supported; a Pearson correlation indicated that the number of psychology courses students had completed was positively correlated with the scores they attained both in January (\( r = .397, p < .001 \)) and in April (\( r = .291, p < .001 \)).

To identify areas of strength and areas for improvement, we performed an analysis of our students’ performance on the seven APA subgoals assessed by our instrument. In order to prepare the data for this analysis, scores on the individual questions were first converted to proportions by dividing by the maximum number of points allotted for each question so that a score of 1.0 was equivalent to obtaining the highest possible score on each question. Next, scores were averaged across pretest and posttest administrations. Finally, questions assessing the same APA subgoal were combined (e.g., Questions 8 and 9 were averaged to assess APA Subgoal 2.2). A one-way repeated measures ANOVA of the seven APA subgoals yielded a significant effect, \( F(6, 348) = 26.38, p < .001, \eta^2 = .31 \). As shown in Table 2, students performed best on Subgoals 2.1 (describing the characteristics of the science of psychology), 2.4 (designing and conducting basic studies), and 2.5 (following the APA Code of Ethics) and worst on Subgoal 2.2 (explaining different methods).

### Table 2. Students’ Mean Performance on APA Guideline Learning Outcomes

<table>
<thead>
<tr>
<th>APA Guideline</th>
<th>Description</th>
<th>M*</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Describing the characteristics of the science of psychology</td>
<td>.77</td>
<td>.31</td>
</tr>
<tr>
<td>2.2</td>
<td>Explaining different methods</td>
<td>.37</td>
<td>.16</td>
</tr>
<tr>
<td>2.3</td>
<td>Evaluating the appropriateness of conclusions derived from psychological research</td>
<td>.55</td>
<td>.19</td>
</tr>
<tr>
<td>2.4</td>
<td>Designing and conducting basic studies</td>
<td>.64</td>
<td>.20</td>
</tr>
<tr>
<td>2.5</td>
<td>Following the APA Code of Ethics</td>
<td>.70</td>
<td>.21</td>
</tr>
<tr>
<td>2.6</td>
<td>Generalizing research conclusions appropriately</td>
<td>.50</td>
<td>.18</td>
</tr>
<tr>
<td>7.1b</td>
<td>Using APA format</td>
<td>.59</td>
<td>.31</td>
</tr>
</tbody>
</table>

*Means converted to proportions ranging from 0 (no credit) to 1 (full credit). bPart of Communication Skills goal and not Research Methods in Psychology.

Although students could potentially have earned a 29 on the assessment, the mean score attained by our RMII students at the end of the term was 17.63 (a score of 58%). One might think that we should be disappointed that our students did not score higher on the test; however, we intentionally included questions that required a high level of sophistication, and our rubric was designed to avoid ceiling effects. We were not concerned about the absolute scores our students attained on the instrument; our goal was to assess improvement over time rather than achieving a specific cutoff score. Hopefully, greater student sophistication would be achieved through subsequent experiences in upper-level and capstone courses.

We took several steps to control for potential confounds. Both of these courses are repeated by some students because they are required courses in which students must attain a minimum grade of a C. We ran analyses to compare the pretest scores of students who were repeating either course to those taking it for the first time and found no difference in the scores for either course. These results were also reassuring because they provided evidence that the students who were repeating the course should repeat it because they had not learned the material the first time.

Each rater graded students from both courses and both administrations and were blind to this information. These ratings were consistent across raters, with excellent interrater reliability coefficients. We did have a training session where we developed and discussed the rubric we used to guide point allocation, which clearly contributed to the reliability levels achieved. The present study points to a strong need for departments embarking on a similar study to spend time on rater training before test administration. This training process also generates productive discussions among instructors about the skills we expect students to attain via our Research Methods courses.

A strength of this technique is its flexibility. If other departments adopted this assessment strategy, the scoring system that we used could be adjusted based on the interests of the specific psychology program. The focus of the questions asked, though, should balance the instructors’ goals with APA Guidelines. It is not essential to use an abbreviated article as we did, but it is critical to use a brief article that is not too statistically sophisticated but will maintain the interest of the students.
Implications

We have used the results of this study in multiple ways. First, it is helping us satisfy external demands from regional accreditation bodies as well as university expectations that we assess student learning outcomes and improve the instructional process with the information that we glean from these analyses. Second, we are using the results to help shape our curriculum. Critically analyzing the results and understanding the weaknesses of the research were the areas in which students underperformed. For example, to justify their arguments for making causal inferences, students consistently relied upon the statistical significance of the results rather than the type of design used. Therefore, we focus on those concepts more in the statistics and methods courses. Given that students would benefit from increased practice with reading and critically analyzing empirical journal articles, we added a capstone course focused on critical analysis of current research. This course is designed to be taken after the Research Methods sequence to help students become more effective and critical consumers of psychological research and ultimately better critical thinkers.

In addition, supplemental analyses of the courses taken by students indicated that certain courses were related to higher pretest scores for RMI, and we are currently exploring whether these courses should serve as prerequisites to the methods sequence to allow students to be better prepared. Third and most important, the study is helping us understand whether, and how well, we are reaching departmental goals. One of our goals, derived from the APA Guidelines for the Undergraduate Psychology Major (APA, 2007), refers to having students "understand and apply basic research methods in psychology, including research design, data analysis, and interpretation" (p. 13). The results of the present study demonstrate that students taking RMII understand and can apply more knowledge about research methods compared to students in RMI and that both groups of students gain substantial knowledge of research methods over the course of a semester.

Acknowledgments

We would like to acknowledge the anonymous suggestions of three reviewers and thank them for improving this manuscript.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interests with respect to their authorship or the publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research and/or authorship of this article: This research was supported by an assessment grant from Eastern Connecticut State University.

References


Bachiochi et al.


### Note

Send correspondence to Matthew R. Kelley, Department of Psychology, Lake Forest College, Lake Forest, IL 60045; e-mail: kelley@lfc.edu.

---

**Research Methods Courses and the Scientist and Practitioner Interests of Psychology Majors**

Karen Manning, Peter Zachar, Glen E. Ray, and Steven LoBello

*Auburn University Montgomery*

This study examined the effects that exposure to research methodology coursework has on students’ interests in scientist and practitioner activities. Consistent with previous research, there was a positive correlation between scientific and practitioner interests. Exposure to instruction in research methods was associated with a loss of interest in scientific activities even for students who had strong interests in scientific occupations.

Interests in research among psychologists, psychology graduate students, and undergraduates is a much-discussed issue (Kimble, 1984; Mallinckrodt, Gelso, & Royalty, 1990; Stricker, 1997; Vittengl et al., 2004). Leong and Zachar (1991) and Aspenson et al. (1993) demonstrated that interests in research and interests in practice are negatively correlated among graduate students. With undergraduate psychology majors, a somewhat different picture has appeared. Both Zachar and Leong (1992) and Leong, Conant, and Zachar (2004) found that among psychology majors, scientist and practitioner interests are positively correlated.

The differences between undergraduate and graduate students raise questions about the occupational interests of psychology majors. We suggest that undergraduate students enter the major with a positive bias toward psychologist activities, a bias that enhances the positive correlation between scientist and practitioner interests. Students presumably form more discriminating views of what does and does not interest them with increased exposure to the discipline.

In this study we examined the relation between exposure to a research methods class and scientist and practitioner interests. We have observed that many students enter psychology with a lay conception of psychologists as clinicians rather than as researchers and that their lay conceptions of science do not include detailed knowledge of research logic. Based on these observations, we hypothesized that students would lose interest in scientific activities after exposure to a class in research methods where such topics as operationalization and hypothesis testing are a main focus. We also hypothesized that psychology majors with strong interests in investigative occupations such as chemist and biologist would be less likely than other psychology majors to lose interest in research after exposure to a research methods class. Furthermore, we hypothesized that exposure to the content offered in a research methods class would have no effect on practitioner interests.
Participants

Participants were 53 psychology majors taking classes in research methods at two colleges, one in the Midwestern United States and one in the Southeastern United States. Seventy-nine percent of the students were women; 8% were freshmen, 26% sophomores, 37% juniors, and 29% seniors. Ages ranged from 18 to 47 with 68% of the participants being between 18 and 25 years old. All participants had taken fewer than seven psychology courses at the conclusion of the study.

Instruments

We measured scientist and practitioner interests with the Scientist–Practitioner Inventory (SPI; Leong & Zachar, 1991; Zachar & Leong, 1992, 2000). The SPI is a 42-item inventory measuring interests in scientist and practitioner activities. It uses a Likert scale ranging from 1 (very low interest) to 5 (very high interest). The 10-year test–retest reliabilities of the scientist and practitioner scales for a sample of PhD-level psychologists were .50 for the scientist scale and .73 for the practitioner scale. For the current sample, on both the pre- and posttests, the coefficient alphas were .90 or greater for scientist and practitioner interests. The subscale alphas were between .74 and .91 for both pre- and posttests.

To examine changes in interests over the course of a semester, we computed difference scores by subtracting posttest scores from pretest scores. Using a formula that took into account the reliability of each test and the correlations between the tests to estimate the reliability of the difference scores (Kaplan & Saccuzzo, 2005), the reliability of the scientist difference score was .78 and the reliability of the practitioner difference score was .70.

We measured vocational interests with the Vocational Preference Inventory (VPI; Holland, 1985). This 160-item test assesses preferences for various occupations. According to Holland, these preferences are stable over time. The six interest traits of the VPI are realistic, investigative, artistic, social, enterprising, and conventional. For our sample, the Cronbach’s alphas were between .65 and .85.

Procedure

Students taking research methods courses from five different professors participated voluntarily and could withdraw from the study at any time without penalty. Professors introduced the SPI and the VPI on the first day of class. Students completed the SPI in class. They took the VPI home and brought completed questionnaires back to the second class. The posttest included the SPI but not the VPI and took place on either the last day of class or during finals week.

Method

Results

The Pearson correlation between scientist and practitioner interests on the pretest was $r(53) = .57, p = .001$; the correlation between scientist and practitioner interests on the posttest was $r(53) = .54, p = .001$. There were no significant differences between the scientist–practitioner pretest and scientist–practitioner posttest correlations.

We conducted a series of paired samples $t$-tests to examine differences between pretest and posttest scores (see Table 1). Using a Bonferroni correction to control for family-wide error rate, the required significance level was .01. As seen in Table 1, students’ interests in scientist activities, practitioner activities, research activities, and therapy activities were significantly lower at the end of the semester.

To compute the difference scores, we transformed the pre- and posttest scientist interests to standard score units based on the pretest mean and standard deviation and then subtracted posttest from pretest. We did the same for the pre- and posttest practitioner scores. The mean score for the scientist difference score was $.49 + \frac{z}{\sqrt{2}}$ score units, which was significantly different from zero, $t(52) = 4.33, p = .001$, $\eta^2 = .26$. The mean score for the practitioner difference score was $.33 + \frac{z}{\sqrt{2}}$ score units, which was significantly different from zero, $t(52) = 3.05, p = .004$, $\eta^2 = .15$. Each score indicates a loss of interest from the pretest to the posttest.

We ran Pearson correlations between loss of interest in scientist activities and Holland’s (1985) VPI variables. Contrary to our hypothesis, there was no correlation between loss of interest in scientist activities and investigative occupational interests, $r(53) = .17, p = .23$. Students with higher interests in investigative occupations were not less likely to lose interest in scientist activities after exposure to research methods.

Using a median split on the Holland investigative variable, we created a group of higher scoring ($n = 25$) and lower scoring ($n = 28$) students. The scientist difference score for the high-scoring group was $.64 + \frac{z}{\sqrt{2}}$ score units. For the low-scoring group, it was $.37 + \frac{z}{\sqrt{2}}$ score units. There was no significant difference between these groups on the difference score, $t(51) = 1.19, p = .24$. The low investigative group lost interest in scientific activities from pretest ($M =$

| Table 1. Differences Between Pretest and Posttest Scores on Scientist and Practitioner Interests |
|----------------------------------|-------|-----|-------|------|--------|--------|
| Variable            | Pretest M | SD  | Posttest M | SD  | df | t |
| Scientist          | 59.2     | 13.4 | 51.9     | 15.5 | 52 | 4.33* |
| Practitioner        | 75.9     | 13.1 | 71.6     | 13.9 | 52 | 3.05* |
| Research           | 27.3     | 6.3  | 23.9     | 6.8  | 52 | 2.58* |
| Therapy            | 48.8     | 8.9  | 46.2     | 9.3  | 52 | 4.07* |
| Statistics         | 7.6      | 2.3  | 6.8      | 2.7  | 52 | 2.09  |

*Note. N = 53. *$p < .01$. 

Vol. 33, No. 3, 2006
The results indicated that, after exposure to a research methods class, psychology majors lost interest in the scientific activities performed by psychologists. This loss of interest is consistent with our positive bias hypothesis, which states that as students enter the field of psychology, they may have unrealistic self-assessments about what they would like to do were they to become psychologists.

Exposure to research methods classes was also associated with loss of interest in practitioner activities. This unexpected finding is not easy to explain, but may be important. Consistent with our positive bias hypothesis, perhaps it is associated with a more general loss of interest in psychology due to learning about scientific activities from a nuts-and-bolts research design standpoint.

Contrary to our hypotheses, higher levels of interest in investigative occupations such as biologist and chemist did not buffer students from losing interest in psychology-related scientific activities after taking a research methods class. If anything, there was a trend for those with higher general scientific interests to have a greater loss of interest in research activities. Students with general scientific interests, however, were still likely to prefer research activities more than other students.

An alternative explanation for the loss of interest is that the lecture format that these courses used is not the best way to interest psychology majors (mostly women) in research methods and statistics. The laboratory portion of these classes primarily involved writing American Psychological Association-style research proposals rather than doing experiments. It is possible that a hands-on and socially oriented approach involving research teams might produce a different outcome.

It would be helpful to know if the loss of interest in both research and clinical activities after exposure to methodology classes is stable. Our anecdotal observations suggest that continued exposure over time may enhance liking for this kind of academic material in a way that an initial experience does not, at least for some students. Research in social psychology suggests that repeated exposure might be more effective for students who are less bored by the material (Bornstein, Kale, & Cornell, 1990; Zajonc, 1968). Based on these findings, we have begun experimenting with a more hands-on and social approach in our curriculum.

Practically speaking, academic advisors and classroom professors may want to note that even students who express high levels of interest in science may lose interest in research activities following exposure to a formal design course. Programs may want to both build repeated exposure experiences into their curricula and suggest to students that they may come to like the material better as their familiarity with it increases. It is possible that many faculty learned to like research methods in the same way.

Discussion

References


Note

Send correspondence to Peter Zachar, Department of Psychology, Auburn University Montgomery, P. O. Box 24023, Montgomery, AL, 36124; e-mail: pzachar@mail.aum.edu.

196 Teaching of Psychology
Psychology is a Science: At Least Some Students Think So
Jeffrey D. Holmes and Bernard C. Beins
Teaching of Psychology 2009 36: 5
DOI: 10.1080/00986280802529350

The online version of this article can be found at:
http://top.sagepub.com/content/36/1/5

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Society for the Teaching of Psychology

Additional services and information for Teaching of Psychology can be found at:
Email Alerts: http://top.sagepub.com/cgi/alerts
Subscriptions: http://top.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Jan 1, 2009
What is This?
TOPICAL ARTICLES

Psychology Is a Science: At Least Some Students Think So

Jeffrey D. Holmes and Bernard C. Beins
Ithaca College

The American Psychological Association’s (2007) curricular guidelines recommend that students develop both an understanding of how psychologists do research and an appreciation for why scientific thinking is necessary. We surveyed a large sample of psychology majors on specific interests, as well as individual difference variables relevant to scientific thinking. Our results suggest that over time, students’ knowledge of scientific thinking increased, whereas their tendency to see psychology as a science did not. Further, students reported greater interest in practitioner activities than scientific ones, and these divergent interests were associated with differential ways of thinking and of viewing the field of psychology. We discuss some implications for conceptualizing undergraduate instruction given that some student characteristics are more malleable than others.

The American Psychological Association (APA) Task Force on Undergraduate Psychology Major Competencies (APA, 2002) explicitly identified an understanding of research methods and a respect for scientific thinking in their list of critical learning goals for undergraduate psychology. Nonetheless, teachers of psychology often express concern about psychology students’ lack of interest in statistics and research methods. Indeed, it is common to hear instructors complain that students simply do not appreciate the importance of understanding how psychologists know what they know. We share these concerns and have ourselves lamented many students’ lack of interest in research. Like others, we value empirical methods for understanding human behavior and are constantly on the lookout for more effective ways to illustrate their importance. However, we have observed that even what seem to be the most interesting (to us) illustrations of the importance of research using the media and other sources often do little to change some students’ opinions about research. We have observed that students often believe either that (a) systematic observation is unnecessary for understanding human behavior or is at least inferior to personal and anecdotal experience, or (b) studying summaries of research findings is useful, but it is unnecessary to examine methodology.

One question concerns the degree to which psychology students see psychology as a scientific discipline. Friedrich (1996) noted that most undergraduate psychology classes include an emphasis on empirical research, but educators seldom evaluate the degree to which students develop true appreciation for the relevance of science to the study of human behavior. Gardner and Dalsing (1986) found that students come to college with many misconceptions about the field of psychology and that, for many students, these misconceptions remain even after taking several psychology courses. Gardner and Dalsing speculated that students likely learn many false beliefs about psychology from parents and the media. Because the beliefs are part of the “conventional wisdom” (p. 34), they are resistant to change even in the face of specific contradictory information based on research. Given this background, it is unclear to what degree beginning college students choosing psychology as a major are truly aware that they are electing to study a science at all. Some researchers (e.g., Berthold, Hakala, & Goff, 2003) have noted that it is important for students not only to learn...
about research methods, but also to appreciate that psychology is a scientific discipline. Friedrich (1996) recommended that instructors convey to students that psychology is in fact a science and suggested that the tendency to view psychology as a scientific discipline should increase with course work and training. However, few researchers have investigated how students’ views of the field and professional interests change as they progress through their psychology course work.

Such observations led us to a meta-examination that transcended the effectiveness of specific examples and teaching strategies. Rather than looking for new ways to teach research methods, we sought to better understand why many undeniably excellent teaching ideas are effective for teaching students concrete methods but not for changing their opinions concerning the necessity of systematic research. Many authors (e.g., McGovern, Furumoto, Halpern, Kimble, & McKeachie, 1991) have advocated for a strong research focus as part of undergraduate curricula. Although this advocacy is clearly warranted, approaches often appear to be rooted in the assumption that students are blank slates and that good instruction in research methods will not only show students how to do research, but will also convince them that scientific thinking and analysis are important for understanding human behavior. Seldom addressed are the personal characteristics and perspectives students bring with them to their course work that may influence their openness to scientific thinking.

Our concern was less about getting students to understand how psychologists do research than determining why so many students are not interested in understanding how psychologists do research and do not see it as relevant to their study of human behavior. Our goal was to explore the interests and views that psychology students bring with them and develop as they progress through their degree program. We began with the existing framework of the scientist-practitioner model and elaborated by adding further individual difference variables in pursuit of a more comprehensive perspective.

The scientist-practitioner model emerged at the Boulder Conference in 1949 as a framework for training clinical psychologists (Baker & Benjamin, 2000). A primary assumption of the model is that students of psychology can become active and informed consumers of research, even if they never become producers of research. The advantage of this approach was in recognizing that many, if not most, psychologist trainees will never become researchers and that ignoring this fact during training might be counterproductive. The developers of the scientist-practitioner model advocated a sort of pedagogical compromise by suggesting that students receive training to actively and critically apply empirical research findings in their professional activities. Doing so would still require an adequate grasp of empirical techniques, but would make explicit the relevance of research across the spectrum of psychological careers.

One can conceptualize undergraduate psychology training in a similar way. Many psychology majors wish to become counselors and therapists or to pursue other applied fields, but might have little interest in the science of psychology. Malin and Timmreck (1979) reported work in the clinical and counseling fields as by far the most popular career goal among the psychology majors they surveyed, and there is little to suggest that this pattern has changed over the past three decades. Further, students considering what is important for psychology majors to learn prioritize helping people far above scientific principles and evaluating research (McGovern & Hawks, 1986). An ongoing struggle (since the Boulder model first emerged) has been finding ways to get students to appreciate the importance of understanding research methods if they do not ultimately want to do research. Based on past research (e.g., Manning, Zachar, Ray, & LoBello, 2006), we suspected that the typical undergraduate psychology major has much greater interest in practitioner activities than in scientific ones. Instructors seeking to get these students interested in research would encounter the same tribulations faced by anyone attempting to motivate students toward a goal that holds little interest and little perceived value.

The idea that students studying psychology may have vast differences in their personal and professional orientations is not new. Zachar and Leong (1992) found that students with greater scientist interests tended to have an objectivist orientation and tended to “endorse impersonal causality and emphasize behavioral contents, elementarism, physicalism, and quantitative analysis” (p. 670); in contrast, students with greater practitioner interests had a subjectivist orientation and tended to “score high on belief in personal will, emphasize experiential contents, are holistic and nonphysicalistic, and endorse qualitative analysis” (p. 670).

Importantly, Zachar and Leong (1992) also used Holland’s (1966) framework for describing vocational interests and found that scientist interests were positively associated with the investigative code and negatively associated with the social code. Practitioner interests showed the exact opposite relation,
being associated negatively with investigative and positively with social. The researchers concluded that the Boulder model for training students to be scientist-practitioners is unrealistic because students’ interests and views of psychology are strongly influenced by personality and that a scientific approach is inherently inconsistent with the personalities of many psychology students. Zachar and Leong (1992) doubted that education could change this pattern, given the stability of personality traits. They argued that trying to force students with practitioner orientations to be scientists “makes as much sense as trying to convert an introvert into an extrovert” (p. 676). In fact, subsequent research suggested that psychology majors’ interest in scientist activities actually declined after taking a research methods course (Manning et al., 2006).

Following up on their earlier research, Zachar and Leong (2000) surveyed their participants 10 years after the original study. At the time of the initial study, the participants had been doctoral students in clinical, counseling, and experimental programs; by the time of the longitudinal follow-up they had completed their graduate work, and Zachar and Leong surveyed them about their professional activities. The investigators reported that scientist and practitioner interests were negatively correlated both when the participants had enrolled in graduate school and also 10 years later. They also found that both types of interests predicted similar interests and professional activities 10 years later. Further, of various professional activities, clinical and counseling psychologists reported the least interest in statistics and research. Zachar and Leong (2000) concluded that despite stated objectives, most graduate students in their sample who were trained under the Boulder model were not scientist-practitioners.

Hypotheses

Whereas previous researchers have studied the views of graduate students, we extended such work by applying the scientist-practitioner framework to better understand the interests and motives of undergraduate students. Based on our objectives and on past research, we proposed several hypotheses. First, we predicted that psychology majors overall would report greater interest in practitioner than in scientist activities. Second, based on Zachar and Leong’s (1992) assertion that scientist and practitioner interests reflect personality characteristics, we expected that mean scientist and practitioner interest levels would remain stable across levels of the curriculum. Third, we expected that scientific literacy would increase as students completed more psychology course work. Finally, based on Friedrich’s (1996) report, we expected that students who had completed more course work in psychology would have a greater tendency to view psychology as a scientific discipline than students who had completed fewer courses.

Method

Participants

The sample consisted of 201 students at a private liberal arts college in the northeastern United States, of whom 147 (73%) were women and 54 (27%) were men. Participants were psychology majors from courses at four levels of our research curriculum starting at the introductory level with General Psychology Laboratory (n = 68), followed by Statistics (n = 29), then Research Methods (n = 44), and finally Research Team (n = 60), which is a three-semester requirement for psychology majors. Students completed the surveys at the beginning of the semester before they had completed any curricular work for that semester. This timing guaranteed that first-year students would have had the least possible exposure to psychology as a science.

Instruments

Scientist-practitioner interests. The Scientist-Practitioner Inventory (Leong & Zachar, 1991) contains 21 items designed to assess scientist interests (e.g., Designing an experiment to study a psychological process) and 21 items to assess practitioner interests (e.g., Conducting a psychotherapy session with an individual client) of psychology students. Respondents rated their interest in each of 42 professional activities on a 5-point scale ranging from 1 (very low interest) to 5 (very high interest). Leong and Zachar (1991) reported an alpha coefficient of .91 for the Scientist subscale and .94 for the Practitioner subscale in a sample of undergraduate psychology majors; the alphas for the current sample were .91 for the Scientist subscale and .88 for the Practitioner scale.

Views of psychology as a science. The Psychology as Science scale (Friedrich, 1996) contains 15 items measuring the degree to which respondents view psychology as a science. Respondents rate items such as, “Research conducted in controlled laboratory settings
is essential for understanding everyday behavior” on a 7-point scale ranging from 1 = (strongly disagree) to 7 (strongly agree). Friedrich (1996) reported a coefficient alpha of .71 for a sample of undergraduate psychology majors. The alpha for the current sample was .72.

**Need for cognition.** The Need for Cognition Scale (Cacioppo, Petty, & Kao, 1984) assesses the “tendency to engage in and enjoy effortful cognitive endeavors” (p. 306). The scale contains 18 items (e.g., “I would prefer complex to simple problems”; “I only think as hard as I have to”) that respondents rate on a 9-point scale ranging from 1 (very strongly disagree) to 9 (very strongly agree). The authors reported an alpha coefficient of .86; the alpha for the current sample was .88.

**Scientific literacy.** A test of scientific literacy (Carrier, 2001) tested participants’ understanding of the scientific process. The test contains 24 true–false items such as, “A scientific law will not change because it has been proven true.” This instrument has unknown psychometric properties but has a high degree of face validity.

**Results**

Students reported significantly higher levels of interest in practitioner activities than in scientist activities across all four curriculum levels. For the sample as a whole, a paired-samples t test indicated that the mean for practitioner interests (M = 74.97, SD = 12.71) was significantly greater than the mean for scientist interests (M = 57.19, SD = 14.39), t(200) = −13.37, p < .001, r = .55. Leong and Zachar (1991) found that scientist and practitioner interests were positively correlated among undergraduates but negatively correlated among psychology graduate students. The correlation coefficients between the two scales for our sample were .19, .14, .10, and −.11 for students at the General Psychology, Statistics, Research Methods, and Research Team levels, respectively. The coefficient for students at the Research Team level differed significantly from the coefficient for those at the General Psychology level, Fisher’s z = 1.7, p < .05, one-tailed. There were no mean differences across the four levels on either of the interest scales, but the shift in correlations may reflect a refinement of interests as students progress through the curriculum. Consistent with past findings, the trend was for the coefficients to shift from positive to negative as students became more advanced. Our students in advanced research courses were similar to graduate students in their pattern of scientist and practitioner interests.

There were significant differences across the curricular levels on scientific literacy, with levels significantly higher among students at the Research Team level than among those at the General Psychology level, F(3, 197) = 2.88, p < .05, η² = .04. Overall, the total number of psychology courses taken correlated positively with scientific literacy, r(198) = .17, p < .05. Importantly, scientific literacy is also the only scale we administered that represented factual knowledge about the process of science rather than an attitude or personality construct. Further, scientific literacy did not correlate significantly with scientific interest or Psychology as Science scores at any level of the curriculum.

Importantly, scientific interests across the entire sample were positively correlated with need for cognition, r(196) = .49, p < .001, and Psychology as Science scores, r(197) = .53, p < .001. Neither of these latter two variables correlated significantly with practitioner interests. Our prediction that students would increasingly come to see psychology as a science as they progressed through the psychology curriculum received no support. There were no mean differences in scores on the Psychology as Science scale across the four research curriculum levels. Further, the zero-order correlation between Psychology as Science scores and the total number of psychology courses taken was non-significant, r(198) = .09, p = .19.

**Discussion**

Our findings support Zachar and Leong’s (1992) contention that students’ interests concerning professional activities in psychology are relatively stable and reflect personality traits. Scientific interest did not correlate with scientific literacy at any level. That is, there was no association between an understanding of the process of scientific inquiry and self-reported interest in scientific activities. One might expect this finding given that scientific literacy reflects factual knowledge that one can learn over time regardless of one’s personal views, whereas scientific interest is an individual difference variable that is far more subjective. This finding is consistent with Friedrich’s (1996) concern that, although it is easy to assess students’ knowledge of research design, doing so provides few data to indicate
that students appreciate the relevance of such activities to the field of psychology. Our data suggest that it is possible for students to grasp the principles of research procedures without valuing these procedures as a necessary mode of understanding human behavior. Clearly, efforts to change knowledge are likely to be more fruitful than efforts to change individual personality characteristics that determine how students look at the world. The most brilliant teaching demonstration to illustrate a research concept may be effective for teaching basic methods, but instructors need to devote more effort to strategies that might also increase student appreciation for, if not enjoyment of, research.

With regard to the association between scientist and practitioner interests, our results refine those reported by Leong and Zachar (1991). They found that scientist and practitioner interests were positively correlated among undergraduates but negatively correlated among graduate students. Leong and Zachar speculated that many students come to college knowing few details about psychology as a discipline but with an interest in the field in general. They suggested that the association between scientist and practitioner interests shifts as students learn more about the field and develop more concrete interests in specific pursuits. Our data corroborated this pattern with greater specificity, revealing a gradual trend from positive to negative correlations as students completed more psychology course work. Consistent with Leong and Zachar's findings, the correlation between the scales for students at the introductory level was positive and significant. Although the correlation was of lesser magnitude than that reported by Leong and Zachar for a sample of psychology majors enrolled in introductory psychology, this change could be attributable to the increased presence of psychology courses at the high school level. Although the possibility is clearly speculative and requires further study, it might be that the refinement of interests alluded to by Leong and Zachar is now more likely to begin before students arrive at college. The pattern is clear, however: As students take more psychology courses throughout their undergraduate and graduate years, their interests tend to become more polarized and they become interested either in scientific or applied activities—but generally not both. Further, Leong and Zachar's work suggests that students who find that they are better suited to practitioner pursuits often come to increasingly dislike scientific and statistical pursuits.

In our study, students' reported interests were also consistent with how they viewed their chosen field of study. Students who had higher scientific interests tended to see psychology as a science, but students with high practitioner interests had no consistent tendency to do so. In other words, the students who were interested in scientific activities were the ones who reliably saw psychology as a science. Further, self-reported need for cognition was strongly correlated with scientist interests but uncorrelated with practitioner interests. Perhaps those students who were not interested in science developed their interest in psychology before they were familiar enough with the field to know that scientific study is an integral component; alternatively, perhaps they simply maintained established beliefs that personal experience is the best source of knowledge despite training in the scientific method. Some students certainly maintained strong practitioner interests while viewing psychology as a scientific discipline, but these students appeared to be in the minority. However, the nonsignificant correlations between practitioner interests on the one hand and need for cognition and seeing psychology as a science on the other might in fact be reason for encouragement. They demonstrated that practitioner interests are not antithetical to a scientific perspective, only that the two might not be inherently connected.

The APA's (2007) learning guidelines for undergraduate psychology programs include recommendations that students both understand research methodology and apply critical and scientific thinking in examining psychological issues. The first of these goals concerns factual knowledge (i.e., how science is done), but the second is likely to be more affected by motivational influences based on whether students view scientific inquiry as important. It appears that a large proportion of students fail to see the necessity of the scientific method beyond the requirements of their academic course work.

For instructors, a critical objective is getting students to understand both how research is done (where all the accumulated psychological knowledge comes from) and why empirical observation is necessary. The data suggest that instructors are generally more effective at the former and that the difficulty in accomplishing the latter might in part reflect the "what's going to be on the test" mentality that so many educators abhor. In our program, our concern is how to reach those students who are indifferent to research because they do not see psychology as a science and do not believe that empirical methods are necessary to an understanding of human behavior. Our thinking is consistent with the expectancy-value model of motivation, which asserts that a person must believe that he or she can achieve a particular goal and that goal must have perceived value for motivation toward the goal to be effective.
(Feather, 1982). In the context reported here, we suspect that both an interest in science and a perception that science is valuable to psychology are necessary components of motivating students to engage in the hard work of empirical inquiry (whatever form it may take). If either component is absent, motivation seems unlikely.

It is important to examine the appropriateness of the scientific-practitioner model for psychology instruction. Although its developers intended it as a guide for graduate clinical training, many undergraduate instructors and departments apply the philosophy, perhaps without being fully aware of it. The scientist-practitioner model acknowledges that many or even most students of psychology will not become empirical researchers but encourages all students to become effective consumers of research. It implicitly assumes that everyone can and should be proficient at both the scientific and applied areas of psychology. In particular, it ignores the reality that students and professionals will not voluntarily read and critique research if they do not enjoy these endeavors and do not deem them necessary. This expectation is akin to expecting a pacifist to volunteer for military combat duty. The pacifist, based on external demand, might at times perform such duties, but likely will never be highly motivated to perform them effectively or voluntarily. Zachar and Leong (1992) argued that scientist and practitioner interests are rooted in personality and represent fundamentally different ways of looking at the world. Assuming that academic psychologists wish to maintain their values concerning the necessity of science in psychology (a stance that we strongly advocate), they need to do a much better job of separating the scientific study of human behavior from the nonscientific. If instructors continue to insist that all students of psychology have a firm grounding in the empirical backdrop of psychology, they need to find new ways to demonstrate to students that psychology is truly a science.

Implications for Teaching Goals and Strategies

To help students appreciate the relevance of science in psychology regardless of their specific career goals, we suggest that psychology instructors pursue two lofty objectives. First, we suggest that instructors more fully integrate research methods into all topical areas. Students (and sometimes instructors) compartmentalize research as a particular content area that is separate from the rest of the discipline and that students can study or ignore in isolation. This pattern manifests itself in some introductory classes where instructors cover research methods early in the course and subsequently neglect such material. Perlman and McCann (2005) reported that the vast majority of psychology departments offer a formal research course, but that in most other courses a research focus is tangential and many courses with research opportunities are electives. Instructors should demonstrate to students that scientific observation is a process integral to all areas of psychology. Textbook authors have sometimes attempted to emphasize this point, but descriptions of featured studies often appear on alternately colored pages to which students may pay little attention. As educators, we should work toward developing new activities that encourage students to identify problems and pitfalls with various ways of knowing about human behavior. However, any such strategies should take into consideration the ways of thinking that students acquire long before they enter our classrooms. Our data showed no increased tendency for students to see science as necessary to psychology as they completed more courses. It is therefore not clear that increasing research demands would help students to see the field as a science. Clearly it is not enough to simply demand more of the same without taking into consideration students’ expectations and motivations. We suggest working toward an identity shift whereby both students and instructors would consider it incongruous to call oneself a student of psychology while professing a lack of interest in, awareness of, or concern with the ways that psychologists know what they know and where it becomes paradoxical for students to elect the psychology major because they “hate math and science.”

Our second suggestion is consistent with that of Zachar and Leong (2000), who recommended teaching practitioners about the philosophy of science so that they better appreciate it. Zachar and Leong provided compelling evidence that most practitioners of psychology are not going to be consumers of research and suggested, for example, that practitioners learn how to test hypotheses when working with clients. This practice, they stated, would show students that there can be multiple solutions in addition to the one consistent with the practitioner’s expert insight. We believe that taking such an approach with psychology students has the potential to help them develop the habit of thinking like scientists in their professional activities, a laudable goal regardless of students’ ultimate occupational aspirations.

Fortunately there are some strategies already in the literature that instructors might use to promote these objectives. For example, one group of researchers
(Miller, Wozniak, Rust, Miller, & Sleazk, 1996) found that requiring students to write essays that directly contradicted their inaccurate beliefs about psychology promoted a shift toward beliefs more consistent with scientific findings. In other words, students assigned the task of advocating attitudes consistent with psychological research adjusted their attitudes in the direction of the scientific findings. This approach might also be effective for teaching appreciation for research in general. A student assigned the task of advocating the importance of research in psychology, even if such advocacy contradicts his or her initial attitudes, might come to better appreciate science as a consequence of having publicly argued its merits. It is incumbent on all faculty, regardless of the courses they teach, to explore new pedagogical strategies to promote science appreciation among undergraduate students. Brewer et al. (1993) declared, “The fundamental goal of education in psychology, from which all the others follow, is to teach students to think as scientists about behavior” (p. 169). To this declaration we would add the caveat that there is a difference between teaching students how to teach students to think as scientists about behavior and teaching them that thinking like scientists matters.

References


Notes

1. An abbreviated version of this article was presented at the Eastern Conference for the Teaching of Psychology on Staunton, VA, June 2007.

2. Send correspondence to Jeffrey D. Holmes, Department of Psychology, Ithaca College, Ithaca, NY 14850; e-mail: jholmes@ithaca.edu
Learning Might Not Equal Liking: Research Methods Course Changes Knowledge but Not Attitudes
O. J. Sizemore and Gary W. Lewandowski, Jr.
Teaching of Psychology 2009 36: 90
DOI: 10.1080/00986280902739727

The online version of this article can be found at:
http://top.sagepub.com/content/36/2/90

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Society for the Teaching of Psychology

Additional services and information for Teaching of Psychology can be found at:

Email Alerts: http://top.sagepub.com/cgi/alerts

Subscriptions: http://top.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Apr 1, 2009

What is This?
Learning Might Not Equal Liking: Research Methods Course Changes Knowledge but Not Attitudes

O. J. Sizemore
Iona College
Gary W. Lewandowski, Jr.
Monmouth University

Students completed surveys at the beginning and end of a sophomore-level course on research and statistics. We hypothesized that the course would produce advances in knowledge of research and statistics and that those changes would be accompanied by more favorable attitudes toward the subject matter. Results showed that knowledge did increase significantly, but 4 of 6 attitude measures showed no change. Two attitude measures (perceived utility of research and statistics) showed significant declines. These results demonstrate the independence of knowledge and attitudes and show that attitudinal change is not monolithic. We argue that students’ misconceptions about research might underlie the declines in perceived utility of research and statistics.

Research and statistics courses are a mainstay of undergraduate psychology curricula, regardless of the location or type of institution (Perlman & McCann, 1999a, 1999b). Unfortunately, undergraduates who take these kinds of classes often do not hold favorable views of the course content (e.g., Estes, Chandler, Horvath, & Backus, 2003; Murtonen, 2005; Rajecki, Appleby, Williams, Johnson, & Jeschke, 2005; Vittengl et al., 2004). Attempts to identify reasons for the disfavor have revealed a number of correlates. For example, Vittengl et al. (2004) found that “openness to experience” (p.), academic aptitude, and perceived course relevance were all positive indicators of interest in research. Other researchers have found that measured anxiety regarding statistics is inversely associated with self-efficacy or perceived competence (Onwuegbuzie, 2000; Thompson & Smith, 1982).

Moreover, students surveyed by Estes et al. (2003) expressed skepticism about the utility of psychological research, but not biological research, indicating that the students’ epistemological beliefs were field specific.

In contrast to these studies of undergraduates, Perl and Kahn (1983) measured positive attitudes toward research among graduate students. In a nationwide survey of first-year students in clinical psychology graduate programs, the majority expressed interest in conducting research and stated that they planned to engage in research activities. The incongruence between undergraduate and graduate students’ attitudes could be the result of two obvious, but not incompatible, processes. First, the graduate school selection process might exclude students lacking interest or skill in research and statistics. Second, undergraduates might enter college with negative attitudes about research and statistics, but their attitudes improve as a result of increased exposure and competence in those areas.

We have observed anecdotally that colleagues who teach research and statistics acknowledge students’ less-than-enthusiastic approach at the beginning of the course. However, these same colleagues also report that students’ attitudes are much improved at the end of the semester. Our purpose was to empirically examine students’ knowledge and attitudes before and after a research and statistics course.

Past research has touched on the possibility that taking a research methods course can change student interests (Manning, Zachar, Ray, & LoBello, 2006). Specifically, Manning et al. examined student attitudes toward scientist and practitioner models before
and after taking a research methods course to determine if coursework changed interest in scientific activities. Results indicated that students held significantly lower scientist, practitioner, research, and therapy interests at posttest compared to pretest. However, those authors reported no differences for interest in statistics, suggesting some independence between global versus course-specific attitudes. In contrast, Harlow, Burkholder, and Morrow (2002) found that students enrolled in a quantitative statistics course showed positive changes in both knowledge of and attitudes toward statistics.

Our study extends previous findings by measuring knowledge in research methodology and statistics, as well as directly assessing attitudes regarding research and statistics. We used a pre–post design to assess seven sections of a research methods course offered across three semesters and taught by three different instructors. This sophomore-level course (with laboratory) serves as an introduction to both research methods and statistics and prepares students for junior-level classes in experimental design and psychological statistics. Inclusion of both research methods and statistics training within one course provided us the unique opportunity to assess simultaneous changes in attitudes and skills relative to methods and statistics. We hypothesized that coursework in research methods and statistics would increase student knowledge. Further, we anticipated that exposure to the subject matter and positive changes in knowledge would be accompanied by positive changes in students’ general attitudes and perceptions of ability and utility.

**Method**

**Participants**

Participants were 139 undergraduate psychology majors (110 women, 24 men, 5 not reported) enrolled in a 200-level research methods course at a private university in the Northeast. The participants ranged in age from 19 to 44 years ($M = 21.2$ years). Ninety-five percent of participants were between the ages of 19 and 24. Students’ self-reported college ranks were 48 sophomores, 72 juniors, and 9 seniors (10 not reported). All had previously taken an introductory psychology course; 82% reported completing 11 or more credits in psychology; 44% reported completing 21 or more credits in psychology.

**Materials**

The materials consisted of a survey packet divided into three sections: demographics, attitudes, and knowledge. The demographics section queried respondents about age, sex, race, scholastic status, transfer credits, and psychology courses completed. We developed the attitude survey for this study with a structure similar to the Survey of Attitudes Toward Statistics (SATS) developed by Schau, Stevens, Dauphinee, and Del Vecchio (1995). Our survey consisted of 30 positively and negatively keyed items, concentrated in six areas: attitudes toward research (e.g., Reading articles about research in psychology is something that I enjoy; Cronbach alpha $[\alpha] = .74$), attitudes toward statistics (e.g., It makes me nervous to even think about statistics; $[\alpha] = .89$), perceived utility of research (e.g., The concepts learned in a research class will be helpful to me in the future; $[\alpha] = .91$), perceived utility of statistics (e.g., Taking a statistics course is a valuable part of one’s education; $[\alpha] = .84$), perceived ability in research (e.g., I am good at writing papers dealing with research; $[\alpha] = .78$), and perceived ability in statistics (e.g., Statistics is easy for me; $[\alpha] = .61$). Respondents indicated their attitudes on a scale from 1 (strongly agree) to 6 (strongly disagree). The knowledge section consisted of 15 multiple-choice items that sampled an array of methodological topics (10 items) and statistical topics (5 items) covered in the course. All packets began with the demographics section but we randomized the order of the attitude and knowledge subsections across packets.

**Procedure**

Research assistants surveyed the participants on the first and last day of each course. Course instructors introduced the research assistants to the students in the classes and then left the classroom; instructors were not present during data collection. The research assistants told the students that participation was strictly voluntary, that they had the right to withdraw at any time, and that neither the researchers nor the instructors would have knowledge of who participated and who did not.

To maintain confidentiality yet retain the ability to match individual data for the pre–post analyses, we assigned a unique identification number to each participant. The identifier included the participant’s birth date, sex, and the last two digits of his or her university identification number.
Results

We excluded participants from the data analysis if they were repeating the class (8 students were identified) or if there was no matching participant number for both the pre- and postassessment. The number of pairs available for analysis was 92. For the attitude measures, we used the average for each of the six subscales; knowledge scores consisted of the total correct items.

To assess the role of participant attrition, we compared pretest data from the 92 participants available for analysis with the 33 sets of unmatched pretest data (14 participants provided posttest data that we could not match with any pretest data). Multiple independent-sample t tests revealed no significant differences for any of the dependent variables. These analyses suggest there were no systematic differences between students with complete data and those with pretest data only.

Change in Knowledge and Attitudes

We used multiple dependent-sample t tests to analyze the 92 participants with both pre- and posttest data (shown in Table 1). Participants tested in the last semester of the study did not receive the knowledge survey. As a result, the sample size for analyses involving the knowledge score differs from those for the attitude measures.

Table 1. Differences Between Pretest and Posttest Measures of Attitudes and Knowledge

<table>
<thead>
<tr>
<th>Variable</th>
<th>Pretest</th>
<th>Posttest</th>
<th>df</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Attitude–research</td>
<td>3.64</td>
<td>0.85</td>
<td>3.62</td>
<td>1.05</td>
</tr>
<tr>
<td>Attitude–statistics</td>
<td>3.41</td>
<td>1.12</td>
<td>3.37</td>
<td>1.30</td>
</tr>
<tr>
<td>Perceived utility–research</td>
<td>4.10</td>
<td>1.21</td>
<td>3.90</td>
<td>1.08</td>
</tr>
<tr>
<td>Perceived utility–statistics</td>
<td>3.92</td>
<td>1.03</td>
<td>3.74</td>
<td>1.05</td>
</tr>
<tr>
<td>Perceived ability–research</td>
<td>3.75</td>
<td>0.79</td>
<td>3.82</td>
<td>0.95</td>
</tr>
<tr>
<td>Perceived ability–statistics</td>
<td>3.68</td>
<td>0.78</td>
<td>3.71</td>
<td>0.87</td>
</tr>
<tr>
<td>Knowledge of methods</td>
<td>6.60</td>
<td>2.44</td>
<td>9.40</td>
<td>2.85</td>
</tr>
</tbody>
</table>

Note. N = 91–92 for the attitude measures; N = 72 for the knowledge of methods.

*Opposite of hypothesized direction.

* p < .05. ** p < .001.

As predicted, the course in research methods produced a significant increase in the students' knowledge of research methods. The mean number of correctly answered questions (of a possible 15) rose from 6.60 (SD = 2.44) to 9.40 (SD = 2.85), with t(71) = 9.53, p < .001, r = .75. Separate analyses of the methods items versus the statistical items yielded significant differences in each subcategory.

Contrary to our predictions, none of the attitude or perception scales showed a significant rise in concert with the change in knowledge. Moreover, both scales pertaining to the perceived utility of research and the perceived utility of statistics showed a significant decline, with t(91) = −2.39, p < .05 and t(90) = −2.15, p < .05, respectively. Calculated effect size correlations for the declines in perceived utility of research and perceived utility of statistics were 0.24 and 0.22, respectively.

Correlational Analyses

In an attempt to better understand the observed outcomes, we considered the correlations among the subscales, as well as the knowledge survey, and the difference scores for the two utility measures (see Table 2). Generally, the six attitude measures were significantly correlated with one another. The exception was the attitude toward statistics subscale, which was significantly correlated with only one other subscale, perceived ability in statistics. Unexpectedly, correlation coefficients between the knowledge of methods scores and the attitude subscales were negative. None of the correlations with knowledge of methods was significant except the aforementioned perceived ability in statistics. Collectively, these correlations showed independence between the students’ attitudes and knowledge at the beginning of the study, particularly in the area of statistics.

The declines in utility of research and statistics were significantly and positively correlated with one another, and significantly but negatively correlated with the pretest scores for perceived utility of research and perceived utility of statistics. This pattern of correlations suggests that, for the most part, students did not draw a distinction between research and statistics, perhaps out of a lack of understanding or experience. With one exception, the remaining pretest scores were not significantly correlated with the declines in utility, nor was there a clear pattern of negative or positive correlation.
Multiple Regression Analyses

To help elucidate the results, we undertook multiple regression analyses. We chose five potential predictors from the pretest data set (attitude toward research, attitude toward statistics, perceived ability in research, perceived ability in statistics, knowledge of methods) for inclusion in these analyses. The decline in perceived utility of research and the decline in perceived utility of statistics served as the criterion variable in two separate analyses. Outcomes of those analyses appear in Table 3.

Four of the five variables chosen for the regression analysis were significantly related to the decline in perceived utility of research, with knowledge of methods as the exception. Two of the predictor variables (attitude toward research and perceived ability in statistics) loaded negatively with the decline. In other words, participants with more positive self-reported pretest attitudes toward research and participants with greater self-reported pretest ability in statistics were more likely to have declines in perceived utility of research. Pretest attitude toward statistics and perceived ability in research were positively and significantly related to the decline. Like the correlations, these results do not provide a clear pattern, perhaps because they accounted for little of the variance in the observed decline.

An identical pattern of positive and negative pretest predictors emerged in our analysis of the decline in perceived utility of statistics. In this analysis, however, only perceived ability in research and perceived ability in statistics were significantly related. Again, the predictors accounted for little of the overall variance.

Discussion

As predicted, the results showed that completion of a course in research methods and statistics produced a significant change in knowledge of those areas. Contrary to our prediction, and contrary to the subjective evaluations of many instructors, attitudes toward research and statistics did not improve from pretest to posttest. Instead, students perceived research and statistics as having less utility, not more. Further, and also unexpectedly, attitudes were largely independent of knowledge—the only significant relation was a negative correlation with perceived ability in statistics. The independence of students’ attitudes and
knowledge at pretest might represent a lack of understanding about the nature of these endeavors; the decline in utility suggests that experience might have dampened an unrealistically optimistic perspective.

Following from Onwuegbuzie (2000) and Harlow et al. (2002), it seemed possible that an increase in competence would engender a positive attitude shift, especially in terms of perceived ability. It is surprising, if not disheartening to us as educators, that following a course in research methods, students did not appreciate that their knowledge and skill had improved. Perhaps the perceived usefulness of research and statistics declined because students did not appreciate that they had learned or they had learned something they did not appreciate.

Manning et al. (2006) speculated that changes in the structure and content of their research methods course might have produced more favorable changes in participants’ perceptions. Specifically, they suggested that if the laboratory portion of the methods course were more oriented to conducting research, not simply writing about it, students might have more favorable views. Our findings tend to cast doubt on those assertions. The course examined in this study represents a hybrid format in which students learn about research and statistics through laboratory demonstrations. In addition, students formed teams to develop a research project, collect data, analyze the results, and prepare a manuscript. This course format is consistent with the approach that Manning et al. anticipated might improve attitudes. Moreover, it is reasonable to expect that the format of this course would lead to greater increases in perceived utility compared to other approaches, but such was not the case.

Our findings, and those of Manning et al. (2006), contrast with those of Harlow et al. (2002). In the latter study, students’ gains in knowledge were accompanied by positive changes in attitude (specifically, increases in self-efficacy and decreases in anxiety). It is important to note that their quantitative methods course provided repeated opportunities for individualized attention, clarification of the course material, and feedback regarding each student’s performance. Under those conditions, students might be more likely to recognize increases in their knowledge (self-efficacy), and such recognition might engender a decline in anxiety. Interestingly, Harlow et al. found that precourse skill was related to precourse attitudes, but not postcourse attitudes. Harlow et al. did not provide an analysis of the relation between postcourse attitudes and skill, nor was there any indication of the students’ attitudes toward the subject matter.

The most troubling aspect of this study is the decrease in the perceived utility of research and statistics. These results are consistent with the findings of Manning et al. (2006) that courses in research methods can lead to a decline in students’ interest in both the clinical and experimental activities of psychologists. However, our study is the first to report a coinciding decline in the perceived utility of research and statistics. Manning et al. contended that these negative changes are a function of students’ “unrealistic self-assessments about what they want to do were they to become psychologists” (p. 196).

We believe the critical issue may be a lack of understanding of the goals and applications of psychological research (field knowledge) rather than unrealistic self-knowledge. Students, whether majoring in psychology or not, have considerable misconceptions about psychological phenomena (e.g., Chew, 2005; Furnham, 1993). It seems plausible that they have a fundamental misunderstanding of psychology in general and psychological research in particular. For example, Dillinger and Landrum (2002) reported significant declines in a broad range of student attitudes toward psychology after completing an informational class about the psychology major. We propose that something similar might occur with research and statistics.

For a majority of students, the term research might equate with library research, not empirical research, and the reported pretest attitudes in our study might derive from the former. During the course of a semester, students could become discouraged by the sometimes tedious and often complex nature of empirical research. We also speculate that many students major in psychology as a means of helping people in an applied setting. Students might have reservations about a course in research and statistics, but perhaps believe the course material will complement their aspirations of helping people. Unfortunately (for those students), courses in research and statistics do not commonly focus on using these skills and techniques in the context of direct helping. Thus, at posttest, students view their new knowledge about research and statistics as having significantly less utility. It is an open question whether a more applied focus in research courses or course enhancements, such as those developed by Harlow et al. (2002), could serve as a counterweight to these declines.

In sum, this study shows that a research and statistics course increased students’ knowledge from pretest to posttest, but for the most part left attitudes unchanged. When attitudes regarding perceived utility did change, they were in an undesirable direction. For this reason,
future research should more thoroughly examine the source of students’ initial attitudes toward, and knowledge of, research practices and determine whether attitudes and knowledge change beyond the first course experience. Ultimately, researchers might consider whether alternative approaches to teaching this subject matter will lead to the positive attitude change many educators anecdotally report.

References


Notes

1. We thank the following individuals for their help on this project: Annie Chirchirillo, Lauren Korcz, Natalie Nardone, Alanna Raines, Tara Scarponi, Ryan Sisko, and Kathryn Yeskel.

2. Send correspondence to O. J. Sizemore, Department of Psychology, Iona College, New Rochelle, NY 10801; e-mail: osizemore@iona.edu.