EDITORIAL BOARD

Editor-in-Chief
Dr. Russell Herman, University of North Carolina Wilmington

Editorial Board
Timothy Ballard, Biology  John Fischetti, Education
Caroline Clements, Psychology  Russell Herman, Mathematics and Physics
Pamela Evers, Business and Law  Mahnaz Moallem, Education

Associate Editor
Caroline Clements, UNCW Center for Teaching Excellence, Psychology

Consultants
Librarians - Sue Ann Cody, Rebecca Kemp
Computer Consultant - Shane Baptista

Reviewers
Sharon Ballard, East Carolina University, NC  Scott Imig, UNC Wilmington, NC
Alison Burke, Southern Oregon University, OR  Robert Kenny, University of Central Florida, FL
Barbara Chesler Buckner, Coastal Carolina, SC  Amanda Little, Wisconsin-Stout, WI
Donald Bushman, UNC Wilmington, NC  Gabriel Lugo, UNC Wilmington, NC
Kathleen Cook, Seattle University, WA  Colleen Reilly, UNC Wilmington, NC
Michelle d’Abundo, UNC Wilmington, NC  Jana Robertson, UNC Wilmington, NC
Donald Furst, UNC Wilmington, NC  Tamara Walser, UNC Wilmington, NC
Sarah Ginsberg, Eastern Michigan University, MI  Handoyo Widodo, Politeknik Negeri Jember, ID
Sibylle Gruber, Northern Arizona University, AZ  Jianjun Wang, California State University, CA

Submissions
The Journal of Effective Teaching is published online at http://www.uncw.edu/cte/et/. All submissions should be directed electronically to Dr. Russell Herman, Editor-in-Chief, at jet@uncw.edu. The address for other correspondence is

The Journal of Effective Teaching
C/o Center for Teaching Excellence
University of North Carolina Wilmington
601 S. College Road
Wilmington, NC 28403 USA
FAX 910-962-3427

(ISSN 1935-7869 for limited print issues and ISSN 1935-7850 for the online issues)

©2011 All rights reserved.
CONTENTS

Letter from the Editor-in-Chief
Russell L. Herman ................................................................. 1-3

Scholarship of Teaching

The Impact of Using Randomized Homework Values on Student Learning
Victor Berardi ................................................................. 4-17

What Makes A Good Criminal Justice Professor? A Quantitative Analysis of Student Evaluation Forms
Patrick M. Gerkin and Christopher A. Kierkus ........................................ 18-39

Learning by Doing: An Empirical Study of Active Teaching Techniques
Jana Hackathorn, Erin D. Solomon, Kate L. Blankmeyer, Rachel E. Tennial, and Amy M. Garczynski ..................................................... 40-54

Evaluating and Improving the Assessment of Undergraduate Student Writing In a School of Business at a Large Regional University
David Glew, Tracy Meyer, Becky Sawyer, Pete Schuhmann, and Barry Wray .............. 55-74

Activating Theory in the Introductory Classroom: Erving Goffman Visits Wisteria Lane
Lisa A. Melander and Susan L. Wortmann ..................................................... 75-86

Effective Teaching

Group Work: How to Use Groups Effectively
Alison Burke .............................................................................. 87-95

The Absent Graduate Student: An A-B-A Single-Subject Experiment of Online Discussion Participation
Thomas D. Cox .............................................................................. 96-109

CALL FOR PAPERS

The Journal of Effective Teaching is accepting submissions for review for the Spring 2012 issue. Manuscripts will be due October 31, 2011. The expected publication date will be February 28th. Articles will be accepted in any of the Content Areas supported by the journal.
INFORMATION FOR AUTHORS

The Journal of Effective Teaching is an electronic journal devoted to the exchange of ideas and information about undergraduate and graduate teaching. Articles are solicited for publications which address excellence in teaching at colleges and universities. We invite contributors to share their insights in pedagogy, innovations in teaching and learning, and classroom experiences in the form of a scholarly communication which will be reviewed by experts in teaching scholarship. Articles should appeal to a broad campus readership. Articles which draw upon specific-discipline based research or teaching practices should elaborate on how the teaching practice, research or findings relates across the disciplines. We are particularly interested in topics addressed in the particular Content Areas described at this site, including empirical research on pedagogy, innovations in teaching and learning, and classroom experiences.

The Journal of Effective Teaching will be published online twice a year at the web site http://www.uncw.edu/cte/ET/. All manuscripts for publication should be submitted electronically to the Editor-in-Chief, Dr. Russell Herman, at jet@uncw.edu. Articles will be reviewed by two to three referees.

Manuscripts for publication should:

- Follow APA guidelines (5th Edition).
- Include an abstract and 3-5 keywords.
- Typeset in English using MS Word format and 12 pt Times New Roman
- Articles/essays on effective teaching should be 2000-5000.
- Research articles should be 3000-8000 words.
- Tables and figures should be placed appropriately in the text.

All articles published in The Journal of Effective Teaching will be copyrighted under the Creative Commons "Attribution-Non Commercial-No Derivs" license. The Journal of Effective Teaching will require that the author sign a copyright agreement prior to publication.

<table>
<thead>
<tr>
<th>Deadlines for Upcoming Issues</th>
<th>Spring 2012</th>
<th>Fall 2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>Submissions Due</td>
<td>October 31, 2011</td>
<td>May 31, 2012</td>
</tr>
<tr>
<td>Notification of Acceptance</td>
<td>December 31, 2011</td>
<td>July 31, 2012</td>
</tr>
<tr>
<td>Final Manuscripts Due</td>
<td>January 31, 2012</td>
<td>August 31, 2012</td>
</tr>
</tbody>
</table>
Letter from the Editor-in-Chief

Russell L. Herman
The University of North Carolina Wilmington, Wilmington, NC

This issue marks the fifth year of publication since the Journal of Effective Teaching was revamped in 2006. Since the revamping of the journal, we have had 290 submissions. We have processed over 270 articles and produced eleven issues of The Journal of Effective Teaching with 59 articles ranging over a variety of topics such as designing rubrics in different disciplines, critical thinking, teaching methods, assessment and active learning. These papers have also covered the use of best practices in many common disciplines including some cross disciplinary studies. We are grateful that our authors and others have considered our journal for publishing their papers and look forward to reading future contributions.

The contents of the last three issues are provided below in case readers missed some of the recent papers published in JET. These are accessible at http://www.uncw.edu/cte/et/. The papers in the current issue are listed on the Contents page and an index of earlier papers can be found in the February 2010 issues. We encourage you to look over these listings not only to find interesting articles, but also to identify areas that we have not covered sufficiently. We welcome submissions to the journal at any time of the year.

Vol 10 Issue 1 (February, 2010)

Letter from the Editor-in-Chief: A Pinch of This and a Dash of That - A Recipe for Teaching Excellence
Russell L. Herman ................................. 1
Inquiry Teaching in the College Classroom
Barbara Neuby ........................................ 4
The Role of Intrinsic Goal Orientation, Self-Efficacy, and
E-Learning Practice in Engineering Education
Seung Youn Chyung, Amy J. Moll, and Shelley A. Berg .............................. 22
Teaching Dossier Documents: A Comparison of Importance by Major Stakeholders
Charles A. Burnap, Gary F. Kohut, and Maria G. Yon ............................... 38
She Needs a Haircut and a New Pair of Shoes:
Handling Those Pesky Course Evaluations
Connie Buskist and Jan Hogan ........................................ 51
Toward Enhancing Scholarship of Science Education in College Teaching
Jianjun Wang and Xingkai Luo .................................................. 57
Re-vitalizing the First Year Class through Student Engagement and Discovery Learning
Erin Steuter and Judith Doyle .................................................. 66

1 Author's email: hermanr@uncw.edu

©2011 All rights reserved.
Assessing the Integration of Embedded Metacognitive Strategies in College Subjects for Improved Learning Outcomes: A New Model of Learning Activity
Katherine Pang and Catherine Ross ................................................................. 79

Vol 10 Issue 2 (September, 2010)

Letter from the Editor-in-Chief: What is Critical Thinking?
Russell L. Herman ........................................................................................................ 1-3
Learner-Centered Use of Student Response Systems Improves Performance in Large Class Environments
Samuel B. Pond, III .................................................................................................. 4-17
Teaching the Scientific Method in the Social Sciences
Grace Keyes ............................................................................................................. 18-28
Here’s Your Syllabus, See You Next Week: A Review of the First Day Practices of Outstanding Professors
Barbara A. Iannarelli, Mary Ellen Bardsley, and Chandra J. Foote .................................. 29-41
Online vs. Paper Evaluations of Faculty: When Less is Just as Good
David S. Fike, Denise J. Doyle, and Robert J. Connelly .................................................. 42-54
Successes and Failures Teaching Visual Ethics: A Class Study
Aimee Kendall Roundtree .......................................................................................... 55-61
Fostering Student Engagement: Examining the Roles of Self, History and Cultural Identity
Patricia Comeaux ....................................................................................................... 62-73
“Mind the Gap” in the Classroom
Sarah M. Ginsberg .................................................................................................... 74-80
Promoting Technology-assisted Active Learning in Computer Science Education
Jinzh Gao and Jace Hargis ......................................................................................... 81-93

Vol 11 Issue 1 (February, 2011)

Letter from the Editor-in-Chief: What Makes an Excellent Professor?
Russell L. Herman ........................................................................................................ 1-5
An Exploration of the Value of Service-Learning: Characteristics of Traditional and Honor Service-Learners
Lori Simons, Elizabeth Williams, and Brittany Russell .................................................. 6-18
Statistical Analysis of Cooperative Strategy Compared with Individualistic Strategy: An Application Study
Tawfik A. Saleh ........................................................................................................... 19-27
Using Monopoly to Introduce Concepts of Race and Ethnic Relations
Warren Waren ........................................................................................................... 28-35
Interdisciplinary Team Teaching: An Effective Method to Transform Student Attitudes
Amanda Little and Anne Hoel ..................................................................................... 36-44
Student Engagement and Teacher Practices: Using Hillary Clinton’s Presidential Campaign for Learning Purposes
Sibylle Gruber ............................................................................................................. 45-61
The Use of a Checklist and Qualitative Notebooks for an Interactive Process of Teaching and Learning Qualitative Research
Rebecca K. Frels, Bipin Sharma, Anthony J. Onwuegbuzie, Nancy L. Leech, and Marcella D. Stark ................................................................. 62-79
Cheating In Online Courses: The Student Definition
Deborah A. Raines, Peter Ricci, Susan L. Brown, Terry Eggenberger, Tobin Hindle, and Mara Schiff .............................................................. 80-89
The Impact of Using Randomized Homework Values on Student Learning

Victor Berardi
Kent State University at Stark, North Canton, OH 44720

Abstract

Much of the recent research on homework focuses on using online, web-based, or computerized homework systems. These systems have many reported capabilities and benefits, including the ability to randomize values, which enables multiple attempts by a student or to reduce academic dishonesty. This study reports on the impact of using randomized versus static homework values on student learning as measured by student performance on similar exam questions. Results indicate randomization can have a significant practical and statistical impact on improving exam scores, can reduce the rate of cheating, and can make homework a more accurate measure of student ability as the difference between homework scores and exam scores is reduced.

Keywords: Randomized homework, academic dishonesty, improved learning.

When looking at the history of research into the efficacy of homework for learning, the only thing clear is how muddy the results are. The concern with whether homework is helpful or not to student learning has been researched for decades without resulting in an agreement on its pedagogical value. Cooper (1989) is an important early review paper on the efficacy of homework for improving learning. In well-designed studies that controlled for homework versus no homework, though, only 70 percent showed a positive benefit of homework with high school students receiving more benefit than elementary students. Follow-up studies (Cooper & Valentine, 2001; Cooper, Robinson, & Patall, 2006), meanwhile, have found the grade level effect was still pronounced, that homework generally has a positive or at least nonnegative impact, but that there is still “no strong evidence” (Cooper et al., 2006) between the homework and achievement link. The authors suggest, after all these years, future research is still needed as most research into homework has used overly complex models that cause potential confounding of results (Trautwein & Koller, 2003; Cooper et al., 2006).

While the majority of research into homework focuses on K – 12 students, results at the college level are similarly conflicting. In my area of business education, a frequently cited study by Rayburn and Rayburn (1999) in a managerial accounting course showed consistent improvement on exam performance for those completing quantitative homework compared to those who did not. Peters, Kethley and Bullington (2002), meanwhile, reports that requiring homework in an operations management (OM) course did not im-
prove results on quantitative exam problems and may have had a slight negative effect on overall performance. Peters et al. (2002) surmise that in their study students in the graded homework group adopted a “zero-sum game” mentality at the end of the semester, figuring they had already put in enough of their effort into the course. Furthermore, they note that the exam problems were multiple-choice and may have been too different from the quantitative, problem-solving homework structure to detect performance differences. I found the Peters et al. (2002) results intriguing because I teach OM at a regional campus in the mid-west where in our program OM is an upper-division, core business course taken by all business majors. I had been assigning homework problems in the course, yet too often, performance on exams in my classes would fail to impress.

Interestingly enough, while the case for homework at all levels is still far from clear, much of the homework research since Peters et al. (2002) has introduced yet another factor into the equation: technology. Using or comparing online, web-based and computerized homework systems for quantitative problems to traditional approaches has gained favor. Computer-based systems can be useful for instructors (Heizer, Render, & Watson, 2009), can positively impact student perceptions and efforts in a course (Smolira, 2008), can be particularly impactful for low-motivated students (Peng, 2009), and can help students self-regulate themselves for learning (Hauk & Segalla, 2005). For example, randomized homework values can make copying answers more difficult or allow multiple attempts, where each attempt is different so memorization does not occur. Then, immediate feedback can be given through automatic grading. With capabilities such as these, positive impacts on learning should follow, right?

Unfortunately, conflicting results still appear to be the norm. In math and science, for example, several studies (e.g., Affouf & Walsh, 2007; Zerr, 2007; Burch & Kuo, 2010) find positive impacts of utilizing this homework technology while others (e.g., Bonham, Deardorff, & Beichner, 2003; Allain & Williams, 2006; Demirici, 2007; Kodippili & Senaratne, 2008) do not.

In my area of business education, Palocsay and Stevens (2008) investigates the effectiveness of web-based homework in teaching undergraduate business statistics. While some performance improvements were seen, the variability in instructors and student ability led to the conclusion that homework technique, including data randomization, had little impact.

Heizer, Render, and Watson (2009), on the other hand, report positive impact in using web-based systems in their graduate and undergraduate operations management classes. On homework assignments, scores improved by 17 to 35 percent through multiple attempts using the web-based Prentice-Hall Grade Assist (PHGA) program. In an MBA program, improvement for the group using the PHGA system over the control group as measured by final grade average was nearly 10 points. In addition to improvements in student learning, the authors note gains in efficiency and integrity in grading along with the possibility to lessen academic dishonesty due to randomization of questions and data values.
Rapid development of commercially available homework systems is welcomed even if the jury is still out on how to best use them. For example, like Peters et al. (2002), I was not seeing benefits of using homework on student performance. Unlike Peters et al. (2002), though, my homework problems were very similar to the exams and yet too often students who earned a high score on a homework assignment would suffer a large performance drop, or gap, on similar exam problems. For example, students averaged nearly a 20 point drop on the exam for the same problem type that they had seen on the homework assignment. This large drop occurred even though many students had earned a 100 on the homework. Not only did these results bring into question the value of requiring graded homework, but it pointed to the likelihood that some students were simply copying homework answers.

When I surveyed students about academic dishonesty on assignments as part of the anonymous end-of-semester student evaluation process, though, only a couple students admitted seeing others copy work and none admitted it for themselves. Students may not have been comfortable acknowledging the existence of copying answers but the performance data indicated otherwise.

Indeed, research into academic dishonesty has shown more than 40 percent of students at all institutions acknowledged copying homework assignments and nearly 80 percent acknowledge seeing or participating in some kind of cheating (McCabe & Trevino, 1993). How prevalent academic dishonesty is for business students compared to other college students is a matter of contention, though. A study by Iyer and Eastman (2006) indicates it is lower while McCabe, Butterfield, and Trevino (2006) show a higher rate. Overall, contextual influences on academic dishonesty find factors such as membership in fraternities and sororities, peer pressure, gender, year in school, pressure to achieve, individualization, and perceptions among students about the acceptance and rate of cheating are seen to be important (Davis, Grover, Becker, & McGregor 1992; McCabe & Trevino, 1997; Pulvers & Diekhoff, 1999; Whitley & Keith-Spiegel, 2002).

While understanding these influences is instructive, for the individual teacher, how to effectively minimize dishonesty so true learning can be achieved is arguably most important. In my case, moral appeals and denoting the extent to which students were allowed to discuss the assignments (Whitley & Keith-Spiegel, 2002) were already being made. Furthermore, the assignment values were being changed for each class (Whitley & Keith-Spiegel, 2002) and yet these efforts were being met with limited success. Hence, I was motivated to develop and implement a randomized data approach with the hope it would help. With no commercial systems meeting my exact needs, I developed my own system using capabilities found in Microsoft Office.

In summary, the primary motivation behind the randomized homework approach is to improve learning as measured by student ability to solve similar problems under controlled exam conditions. A major concern was that a large performance gap existed for some students in the course where exam scores on similar problems were much lower than on their homework. This implied that whatever efforts students made or interactions they had with static homework values; it was not being translated into individual per-
formance on exams. As a result, the potential benefit homework could have on learning was not being fully realized. A more targeted design than employed by others (e.g., Peters et al., 2002; Palocsay & Stevens, 2008; Heizer et al., 2009) was developed to avoid confounding results (Cooper et al., 2006). In particular, the design required not only looking at changes in the overall performance on exam problems and homework scores but in matching individual student homework scores directly to their related exam problems. While the study reported in this manuscript has a design that extends the important foundation work of Peters et al. (2002) and Heizer et al. (2009) in the operations management area, it should have applicability to other quantitative, problem-solving courses and be independent of implementation platform.

The remainder of the paper details my efforts to implement and study randomized homework. First, the hypotheses developed for the investigation are discussed, which is followed by experimental design considerations. Then, details on the homework assignments, including generation and grading considerations are covered. Finally, the results are presented and discussed, including limitations of the study and future considerations.

Hypotheses

**H1: Student Exam Performance**

The first objective is to improve student performance on exam questions of the same problem types as found on the homework assignments. This objective is measured via hypothesis H1, which states that student performance on exam problems will improve after implementing randomized homework assignments.

Exams in the course included multiple choice questions, calculation problems with related analysis questions, and essays. The multiple choice questions are almost exclusively qualitative in nature and serve to monitor general concept understanding and to check completion of reading assignments. The majority of point values on exams, however, is associated with calculation problems and related analysis questions.

To measure H1, only the exam calculation problems and analysis questions related to a homework assignment are evaluated. This avoids confounding performance results with concepts and question structures unrelated to homework that potentially impacted results of others (e.g., Peters, et al., 2002).

For example, the homework assignment for statistical process control (SPC) included an $x$-bar and $r$-chart application that had to be identified and calculated by the student from the descriptive information on the assignment. On exam two, a similar set of SPC problems were posed. If these problems represented 60 points on an exam, and a student earned 54 points, an exam performance of 90 percent was recorded for this study. Details on the structure of exam performance determination, especially related to efforts for maintaining consistency of measures between the control group and post-implementation students, are found in the section on experimental design.
It should be noted that the homework, whether randomized or not, always precedes the corresponding exam by at least two class periods. This is intentional so students can review their performance and seek help in advance of the exam. Furthermore, students are encouraged to bring two copies of their submission to class when the homework is due because after collecting their submission I solved it (or a randomized case of it) in class in order to provide some immediate feedback, which has been shown to improve student performance (Yourstone, Kraye, & Albaum, 2008; Smolira, 2008). Graded assignments were returned to students at the next class meeting.

**H2: Student Homework Performance**

While increasing exam performance is the desired effect, the overall performance on the homework assignments is expected to decrease. This represents the second hypothesis, H2. With randomized homework assignments, the process for students to check their work with others in an ad hoc fashion is more complex and requires significant understanding and effort. Students who forget or procrastinate do not have an easy way to simply acquire the correct answers, which should lead to lower homework scores as postulated in H2. Related to this, it was believed that much of the decrease in homework scores would be from students who achieved a perfect homework score before randomization but not because of their own understanding. Therefore, the percentage of students earning a perfect score on homework assignments before and after randomization is monitored and reported.

**H3: Reducing homework versus exam performance differences**

Perhaps the greatest motivation for developing randomized homework assignment values resulted from the observation that there were too many students getting high scores on homework assignments that subsequently performed poorly on the same material on exams. A large gap between a student’s homework and exam performance certainly brings into question the value a homework assignment had for learning. Therefore, in addition to improving the overall exam scores from H1, an important objective is to reduce the individual differences between student exam and homework performance.

The third hypothesis, H3, states that the gap between homework and exam performance on a student-level basis will decrease after implementing randomized homework assignments. In effect, there will be a significant reduction in students scoring very well on homework assignments that are not being able to perform well on similar problems during the exam.

H3 is measured by looking at the difference between each student’s homework score and their exam score on the same problem type. The homework score for each student was readily available as it was recorded directly in the grade book while the student’s exam score was calculated as in H1. The gap between the homework score and the relevant exam performance was then calculated. For example, a student who earned a 100 on homework but only scored 90 percent on the exam material had a difference score of negative 10. H3 differs from the first two hypotheses in that H1 and H2 are measured from an
overall group performance level instead of at the student level and it is expected to have strong identification power.

**Experimental Design Considerations**

Several important issues were considered in the design of this evaluation. These included adopting an appropriate experimental design structure, maintaining assignment integrity, and minimizing grading inconsistencies that could bias results. The efforts addressing these issues are discussed next.

**Experimental Design**

To evaluate the hypotheses, a pre-test/posttest experimental design is employed in this study. The pre-test period represents a control group of 63 students with static homework data while there are 94 students who received randomized values in the implementation phase. In all statistical tests, a conservative assumption of unequal variances was used even when an equal variances assumption could be justified.

There are three homework assignments: forecasting, SPC, and inventory; each corresponding to a different exam through the semester. Forecasting was on exam one, SPC on exam two, and inventory management on exam three (the non-comprehensive final exam). Similar homework assignments and exam questions were used in each phase and special precautions were used to ensure consistency and to maintain assignment integrity as will be discussed below.

**Maintaining Assignment Integrity**

Maintaining assignment integrity is important in any teaching situation but is critical in an experimental study like this. The actual homework assignments were modified each semester regardless of whether randomization occurred or not. The problem types are the same but the actual problems differ and strict control of the questions themselves is not possible, nor attempted, given that students complete homework outside of the classroom and can easily pass copies along. Even when static values were used in the homework, the actual values were changed semester to semester. Exams, though, were controlled carefully to ensure no copies were stolen. Integrity was accomplished by not only counting the copies but in not recording student grades until after the exam was reviewed in class and all exams were verified as returned.

**Minimizing Inconsistencies in Grading**

Another important concern was controlling the consistency of exam and homework grading. Rubrics played an important role here. Effort was made before the control phase began to standardize grading where common mistake areas were identified and point deductions were determined. These rubrics were used for both homework and exam grading to minimize inconsistencies from student to student and over time.
Homework Assignments

This study incorporates three different homework assignments in the randomization experiment. Each assignment, which corresponds to a different exam during the semester, is detailed next. Then, information on assignment generation and the mechanics of grading with randomized data is discussed.

Forecasting

The first homework assignment of the semester relates to forecasting. Students are required to develop forecasts where each student has different data series values and variability. Three forecasting methods are required including moving averages, weighted moving averages, and exponential smoothing. Forecast evaluations are made and error measures are calculated, which include bias, mean absolute deviation (MAD), and mean square error (MSE). Students must ultimately choose the best overall forecasting method, provide an explanation of why it was chosen, and answer additional questions related to the forecasting process they just completed. This is a significant effort for students, especially early in the semester. The concepts themselves are not particularly difficult but the sheer number of new issues is intimidating to students.

Statistical Process Control

The second homework assignment relates to statistical process control (SPC). In this assignment, an organizational application is described and data given. Students must identify the correct chart(s), calculate the control limits, and develop the graphs. They then must answer related questions, such as to identify if the process is in-control or not. In its current form, the application is the same for all students, a variable data situation necessitating x-bar and r-charts, but the values and decisions vary in the randomization phase. For example, one student may have a process that is in-control while others have out-of-control process data.

SPC is typically the most conceptually challenging section of the course for students and considerable time is spent reviewing the probability and statistics theories that underlie the charts. While the textbook is used for theory and understanding, students use The Memory Jogger II (Brassard & Ritter, 1994) for implementation. This pocket reference has a decision tree students use to select the correct chart, a step-by-step guide for control chart implementation, formulas, constants, and information on how to interpret the charts. The Memory Jogger II is a handy reference book that students are allowed to use on exams too. SPC related questions are found on the second exam of the semester.

Inventory Management

The third homework assignment relates to inventory management. As with SPC, a description of an organizational setting is provided to students along with randomized data for costs and usages. Students must identify the correct inventory model, calculate items of interest such as order quantity, reorder point, and total cost, and then answer related
questions. Inventory management is one of the last major topics of the semester and it corresponds to questions on the final exam.

**Homework Generation**

The randomized assignments in this study are implemented using Microsoft Office and were developed by the author. It should be noted that textbook publishers have tools that can accomplish randomization that might suffice for other instructors.

In this implementation, students are sent an Excel file as an attachment to an email or given a file link to download. When the file is opened, only one spreadsheet is visible in the workbook. All other sheets and programming are hidden and password protected. A Visual Basic for Applications (VBA) macro automatically begins, which opens a dialog box prompting the student to enter his/her name and the last four digits of his/her student identification (ID) number. Error checking routines are employed to ensure all four values are entered and that they are indeed numbers and not letters or other characters. Those interested learning more about coding in VBA should find Walkenbach (2007) is an accessible, yet comprehensive, resource.

The ID is used to generate the randomized values for each student where a minimum of 1000 cases are created for each. The student ID is then scrambled before being used to generate data, where a different ID pattern is used for each assignment to make it improbable that two students could have the same or similar values for multiple assignments.

Each case for an assignment has both a systematic adjustment and a random error component. For example, in a forecasting assignment, an initial pattern was developed and a systematic adjustment was used to shift the cases in scale, going from several hundred units in some cases to several thousand units in others. Then, a random adjustment was added to the individual values so the pattern would show variability and not simply be a shift that could be readily identified through graphing.

This approach facilitates the ability to have differing decisions from student to student based upon the specific case data. As a result, not only do the values differ but the analysis and conclusions can vary too. This was explained to students so they would understand it, get over it, and focus on making decisions consistent with what the data indicated was correct for their case without regard, or fear, for what others may have been seeing.

**Grading of Student Assignments**

One of the significant concerns with using randomized assignments for student homework is how to grade efficiently when the values are different for each case. Indeed, at the 2009 Decision Sciences Meeting in New Orleans, several of the publisher offerings for randomized or internet-based homework systems presented in break-out sessions emphasized instructor efficiency given automatic grading capabilities. While the prime motivation in this study is to determine if randomized assignment values do indeed improve
student learning without regard to implementation platform, automatic grading via publisher offerings is an added bonus for instructors with large section sizes.

On my campus, section sizes are kept below 35 students, so maintaining the connection and feedback obtained from personally grading student assignments as part of teaching reflection is still possible. To facilitate the individualized grading key, a solution worksheet was created in the Excel workbook for each assignment. This sheet initially was used for error checking the systems and for reviewing the development of the student cases in each assignment. This worksheet was deleted from the file given to students but it was used in conjunction with the roster and simple looping logic to develop customized solution sheets for each student. These were aggregated into a class key.

The main grading efficiency loss came from the fact that memorizing the correct answers was not possible. However, with good grading key structure, just a little extra time, perhaps 20 to 25 percent, was needed. Of course, online and publisher-based grading tools will further reduce or eliminate efficiency concerns for those interested in using randomized assignment values in their courses.

Results and Discussion

The first hypothesis, H1, tests if student performance on exams improved during the randomized homework phase of the study. In figure 1, it is seen that exam performance increased by 3.83 \( (p\text{-value} = .083) \), 3.63 \( (p\text{-value} = .038) \), and 2.65 points \( (p\text{-value} = .188) \) respectively on each of the three exams. Recall that the first exam covered forecasting, the second included SPC, and the third covered inventory management. Therefore, randomized homework has a statistically significant impact on the first two exams but less of a statistical effect on the third.

From a practical standpoint, it is encouraging to see improvement range from one-quarter to nearly four-tenths of a letter grade for students in the randomized homework group relative to the static homework students. Even from a theoretical standpoint, the gain achieved, as measured by Hake’s gain statistic (Hake, 1998), \( g \), shows that between 11 percent and 17 percent of the total gain possible is achieved simply by implementing randomized homework values. As an instructor, I was pleased with these results, especially considering that during the static data phase, the range of exam averages across the three exams was only 1.37 points, roughly one-half the impact seen from randomization. In other words, randomizing homework values led to a larger impact in student performance than a change in topics—from forecasting to SPC to inventory—had during the static data phase.

It was not initially clear why the forecasting and SPC topics had a greater improvement than the inventory material; however, viewing the results in terms of the end-of-semester gaming phenomenon, where students seek to maximize their grades across courses as posited by Peters et al. (2002), sheds light on the issue. For example, students who have done well in a course and built up a cushion going into the final exam might put more effort into their other courses, resulting in uncharacteristically low scores on the final ex-
am. Alternatively, students who had not performed as well on prior exams, but who still wanted to improve their course grade, could be expected to show relative improvement on the final exam. Fortunately, the data give some insights into this phenomenon.

First consider the exam scores for the students who earned 100 percent on at least one of the first two exams. On average, these students scored 93.8 percent on the first two exams but only 81.5 percent on the final exam. Hence, students who had scored the highest on the earlier exams dropped 12.3 points on the final exam to just a few points above the class average. At the same time, of the students who earned a 100 percent on the final exam, none had earned a 100 on an earlier exam. In fact, these students had averaged only a 78.8 percent on the first two exams, below the average. Both of these results support the proposition that external end-of-semester issues were at play. As an instructor, it is particularly satisfying to see students who had struggled on earlier material rise up to achieve on the final exam.

Figure 1. Data for hypothesis H1: performance on exams will increase with the use of randomized homework assignments.

<table>
<thead>
<tr>
<th></th>
<th>Exam 1 (Forecasting)</th>
<th>Exam 2 (SPC)</th>
<th>Exam 3 (Inventory)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>pre-random</td>
<td>post-random</td>
<td>pre-random</td>
</tr>
<tr>
<td>Mean</td>
<td>77.47</td>
<td>81.30</td>
<td>76.39</td>
</tr>
<tr>
<td>Variance</td>
<td>304.9</td>
<td>255.0</td>
<td>118.0</td>
</tr>
<tr>
<td>Obs.</td>
<td>63</td>
<td>94</td>
<td>63</td>
</tr>
<tr>
<td>t-statistic</td>
<td>-1.39</td>
<td>-1.79</td>
<td>-0.89</td>
</tr>
<tr>
<td>p-value</td>
<td>0.083</td>
<td>0.038</td>
<td>0.188</td>
</tr>
</tbody>
</table>

The second hypothesis, H2, measured student performance on the homework assignments where a performance decrease during the randomization phase is anticipated. In Figure 2, it is seen that homework scores decreased by 2.96 points (p-value = .053) on the first homework assignment, forecasting; was lower by 9.35 points (p-value = .000) on the second assignment, SPC; and yet increased by 2.18 points on the third assignment, inventory. The magnitude of the impact, in terms of Hake’s (1998) statistic, g, is moderate for the forecasting (g = -0.48) and inventory (g = 0.37) and is strong for SPC (g = -1.57).

Figure 2. Data for hypothesis H2: performance on homework assignments will decrease with the use of randomized homework assignments.

<table>
<thead>
<tr>
<th></th>
<th>HW 1 (Forecasting)</th>
<th>HW 2 (SPC)</th>
<th>HW 3 (Inventory)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>pre-random</td>
<td>post-random</td>
<td>pre-random</td>
</tr>
<tr>
<td>Mean</td>
<td>93.89</td>
<td>90.93</td>
<td>94.03</td>
</tr>
<tr>
<td>Variance</td>
<td>89.1</td>
<td>179.8</td>
<td>77.0</td>
</tr>
<tr>
<td>Obs.</td>
<td>63</td>
<td>94</td>
<td>63</td>
</tr>
<tr>
<td>t-statistic</td>
<td>1.63</td>
<td>4.65</td>
<td>-1.24</td>
</tr>
<tr>
<td>p-value</td>
<td>0.053</td>
<td>0.000</td>
<td>0.609</td>
</tr>
</tbody>
</table>

Further review of the data indicates that much of this impact comes from a change in the rate of perfect homework scores after randomization. The rate of perfect scores decreased...
by 15.3 percent on the forecasting assignment, by 49.7 percent on SPC, while perfect scores increased by 12.2 percent on the inventory homework.

To understand the implication of this, consider the impact randomized values had on the scores. For SPC in particular, it is apparent that many of the perfect scores on the homework during the static data phase were the result of copying by students; simply giving each student unique values cut the rate of perfection in half. In the randomization phase, however, this means more students did not have a high score on the SPC homework, artificially propping up their average. As a result, doing well on the final homework assignment, inventory, became more important and students responded. So not only is the apparent rate of cheating decreased, but motivation for students to perform on the next assignment appears increased. While these results are encouraging for the class as a whole, viewing homework’s effect on exam performance from an individual student perspective is more refined and discussed next.

The third hypothesis, H3, tested individual student performance gaps on homework compared to the related exam problems. Figure 3 shows a reduction in the gap between homework and exam performance on all three assignments. On the first exam, forecasting, the performance gap between the static data group and random data group is reduced by 6.8 points, from a gap of -16.4 to -9.6 ($p$-value = .006). On exam two, SPC, the difference is reduced by 13.4 points, from -17.6 to -4.2 points ($p$-value = .000). The third exam, inventory, shows a very small performance gap of 0.5 points, from a gap of -17.9 to -17.4 points ($p$-value = .438). Therefore, on the first two assignments, a significant statistical and practical reduction in individual performance gaps on homework and exam scores is achieved. On the third exam, end-of-semester gaming, as discussed under H1, appears to play a role in the near zero change in performance gap.

**Figure 3. Data for hypothesis H3: individual performance gaps on exams and homework assignments will decrease with the use of randomized homework assignments.**

<table>
<thead>
<tr>
<th></th>
<th>Pre-random</th>
<th>Post-random</th>
<th>Pre-random</th>
<th>Post-random</th>
<th>Pre-random</th>
<th>Post-random</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Forecasting</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>-16.4</td>
<td>-9.6</td>
<td>-17.6</td>
<td>-4.2</td>
<td>-17.9</td>
<td>-17.4</td>
</tr>
<tr>
<td>Variance</td>
<td>236.8</td>
<td>322.4</td>
<td>127.9</td>
<td>301.0</td>
<td>405.8</td>
<td>231.8</td>
</tr>
<tr>
<td>Obs.</td>
<td>63</td>
<td>94</td>
<td>63</td>
<td>92</td>
<td>60</td>
<td>90</td>
</tr>
<tr>
<td>$t$-statistic</td>
<td>-2.53</td>
<td>-5.84</td>
<td>-0.16</td>
<td></td>
<td>0.006</td>
<td></td>
</tr>
<tr>
<td>$p$-value</td>
<td>0.006</td>
<td>0.000</td>
<td>0.438</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The improvements related to the forecasting and SPC material are considerable and warrant added discussion. Notice for the control group, the decline in exam performance on the forecasting and SPC material is 16.4 points and 17.6 points, respectively. This means that a student with static data who gets a homework score in the high 90s could struggle to achieve a low 80s score on the same material under exam conditions. While one could certainly expect to see some decrease given the control and pressure inherent in an exam situation, any instructor would be concerned by this magnitude of effect.

*The Journal of Effective Teaching, Vol. 11, No. 2, 2011, 4-17*  
©2011 All rights reserved
Once again, implementing randomized homework values had a significant practical impact in addition to a statistical one. On the forecasting material, the homework versus exam performance gap was reduced by more than 41 percent to a 9.6 point decline for the students receiving randomized homework values. For SPC, the performance gap was reduced by nearly 76 percent to a satisfyingly small 4.2 point difference. In other words, with randomized data the homework assignments have become better indicators of a student’s ability to perform on the exam. And since overall exam performance increased per H1, student learning, as measured by performance on the same material at a later point in time under controlled conditions, appears improved.

**Conclusion and Future Work**

While the design used in this study was necessarily tailored to the issues as they pertained to homework in an established operations management course, the randomized homework approach and lessons learned should be applicable to other quantitative courses. In this study, exam scores increased when using randomized homework values compared to using static values, thereby supporting hypothesis H1. Student homework scores decreased as postulated in H2, where much of this drop was due to a lower rate of perfect scores on homework after randomization. On one assignment, perfect scores dropped by nearly 50 percent, which indicates that in the static data phase many students were probably just copying answers. Finally, with randomized values, homework scores became a more accurate reflection of student ability as the gap between homework and exam scores was reduced per hypothesis H3.

Randomized values are not a panacea, though, as external factors affecting student performance can be significant. In particular, end-of-semester concerns by students who are trying to maximize their grades across several courses are not trivial. On the positive side, it was seen that the students who perform best on the final exam were those who struggled near the class average earlier in the semester. At the same time, high performing students may still suffer a drop-off as they focus their energies on other courses. Investigating the impact of semester timing on performance by randomizing topic sequencing would be a significant study to conduct in itself. To do so in a meaningful timeframe would require having several sections of a course offered simultaneously so assignment timing can be randomized, which is more than a small campus such as mine affords.

The randomized capabilities developed for this study or those available from publisher offerings may have even greater value for use in fully online courses. In addition to homework, problems presented on exams can include randomized data where values, and indeed problem structure, can be varied for each student. As in this current investigation, there are implications for both improving student learning and reducing academic dishonesty in online sections. This is especially important given that exams in many online courses are completed in an un-proctored environment similar to what is seen with homework from traditional courses. I have just begun teaching a fully online section of operations management and I am collecting data on the initial phase of an investigation.
Acknowledgement

This project received support through a Kent State University Teaching Council’s Teaching Development Award.

References


What Makes a Good Criminal Justice Professor?
A Quantitative Analysis of Student Evaluation Forms

Patrick M. Gerkin¹ and Christopher A. Kierkus
Grand Valley State University, Grand Rapids, MI, 49508

Abstract

The goal of this research is to understand how students define teaching effectiveness. By using multivariate regression analysis of 8,000+ student evaluations of teaching compiled by a School of Criminal Justice at a Midwestern public university, this paper explores the relationships between individual indicators of instructor performance (e.g. clarity of the instructor, concern for student progress etc.), and overall impressions of instructor quality. Further consideration is given to factors like basic student characteristics (e.g. academic status), basic course characteristics (e.g. whether the course is a research methods class), basic instructor characteristics (e.g. instructor gender), and the expected grade in the class. Results indicate that the individual instructor performance items are the most important indicators of overall perceived quality. Other factors do matter, but their influence is statistically weaker. Interestingly, once the individual instructor performance items are controlled, the effect of perceived grade is rendered statistically insignificant.

Keywords: Teaching effectiveness, student evaluations.

Student evaluations of teaching (SETs) appear to be a topic of never-ending debate and inquiry. Research has been accumulating for decades about the value and validity of student’s evaluation of faculty performance in the classroom. This line of inquiry began in 1927 with the first published research on the topic of student ratings (Remmer & Brandenbur, 1927). Student evaluations of teaching at the time were in their infancy as the first record of student evaluations occurred at the University of Washington only a year or two before (Guthrie, 1954). Despite intensive debate over their value, the use of SETs has grown almost continuously at universities across the United States (Marsh, & Roche, 1997; Seldin, 1993).

The ferocity of the debate, and continued scholarly inquiry, are not too surprising given that as the usage of SETs has expanded, so too has their utilization. Student evaluations of teaching are now considered in numerous settings, and at various levels within the university. Today, it is common for SETs to be requested from incoming applicants to be used in the hiring process; they are used at many institutions throughout the tenure and promotion process; and they may also be used as a means of evaluating candidates for teaching awards. Most certainly, faculty use them for formative purposes, to evaluate

1 Corresponding author’s email: gerkinp@gvsu.edu

©2011 All rights reserved
and assess their own performance and to guide their personal efforts to improve their teaching.

While the use of SETs for formative purposes is not particularly controversial, their use as a summative measure, designed to assist in personnel decisions has led to significant objection. Formative evaluations have, in fact, been found to provide useful feedback and may improve undergraduate education (Griffin & Pool, 1998; Marsh & Roche, 1997). The controversy is generally centered around two major issues (1) the validity and reliability of instructional evaluations by students and (2) the objectives and possible applications of such evaluations (Miron & Segal, 1986). Additionally, it should be noted that research conducted by Franklin and Theal (2001) demonstrates a lack of knowledge about the literature pertaining to student ratings, and lack of statistical expertise among the faculty and administrators who are assigned to interpret them. “Even when the data are technically rigorous, one of the major problems is day-to-day practice: student ratings are often misinterpreted, misused, and not accompanied by other information that allows users to make sound decisions” (Franklin & Theal, 2001, p. 46).

The debate over the value, validity and reliability of SETs is further intensified by programs like those implemented at Texas A&M, where according to Katherine Mangan (2009), the school has started awarding bonuses ranging from $2,500 to $10,000 to faculty members who receive the highest SETs at the end of the term. Although critics of this program fear that it will lead to grade inflation as professors strive to win the admiration of their students, Mangan (2009) notes that a similar program was already in place at the neighboring University of Oklahoma. She goes on to describe how another Texas institution, the Acton School of Business, ties instructors pay even more closely with student evaluations. The faculty at this small private institution are paid $5,000 per course, with the remainder of their salary determined solely by evaluations.

As the debate over the value and proper use of student evaluations continues, SETs continue to be utilized for a variety of purposes. This study seeks to make a significant contribution to the scholarship of teaching and learning by offering findings from a large sample of student evaluations compiled by a school of criminal justice at a major university, to determine what variables are most important to students when evaluating the overall quality of their teachers. While there has been much research focused on the validity and potential biases in student evaluations of teaching, far less attention has been given to the significance of individual items in a SET, and their relationship to students’ overall perceptions of teacher effectiveness.

**Literature Review**

No matter their utilization, or actual value, SETs are clearly having an impact on faculty performance in the classroom. One survey of faculty found 70% of professors believe that their grading leniency, and course difficulty, bias student ratings, and 83% admitted making their courses easier in response to student evaluations (Ryan, Anderson & Birchler, 1980). This suggests whether one believes that SETs are valuable, for formative or summative purposes, their use clearly influences faculty behavior in the classroom.
The research included here will provide an overview of significant issues related to SETs, including research devoted to the issues of validity and reliability, instructor attributes such as race and gender, the influence of grades, course characteristics such as class size and type of course, and instructor qualities like clarity, concern for student progress, and fairness.

**Validity**

Hundreds of articles have been written on the validity of SET ratings, and it seems that most researchers agree that validating SETs is no easy task. Marsh and Roche (1997) note that SETs are difficult to validate because there is no single criterion of effective teaching. Kulik (2001) agrees:

> If ratings are valid, students will give good ratings to effective teachers and poor ratings to ineffective ones. The size of the correlation between ratings and effectiveness will provide a precise index of the validity of student ratings. The catch is that no one knows what measure to use as the criterion of teaching effectiveness. Researchers have long searched for the perfect criterion (p. 10).

Despite a failure to find the *perfect* criterion, several criteria have emerged, and appear frequently in the literature. These include measures of student learning, alumni assessments of teachers, and classroom observations by experts. Even these criteria appear less than perfect. Scriven (1983) offers a clear articulation of the inadequacies of these measures.

Given the obstacles identified here, numerous researchers have turned to an alternative construct-validation approach that seeks to relate SET ratings with a wide variety of indicators of effective teaching (Howard, Conway, & Maxwell, 1985; Kulik, 2001; Marsh, 1987). A list of the common indicators includes those previously identified as less than perfect measures when used individually. The idea is that together, they can offer a validation of SETs.

Drawing from the expansive literature on student evaluations of teaching, it is possible to offer a few general conclusions about which there appears to be some consensus. Kulik’s (2001) meta-analysis of the validity and utility of student ratings concludes that “student ratings agree well with other measures of teaching effectiveness; learning measures, student comments, expert observations, and alumni ratings” (p. 23). Moreover, multisection validity studies have been able to demonstrate that SET ratings do reflect the level of student learning; which is an important criterion of effective teaching. Multisection validity studies collect data from multiple sections of the same course, taught by different professors, who administer the same standardized final exam. Two meta-analyses of such studies (Marsh, 1987; Marsh & Dunkin, 1992) suggest that the sections with the highest SETs are also the sections that perform best on the standardized final examination. Cohen (1981), in a widely cited article on the relationship between SET ratings and learning, found significant correlations between ratings and student performance on a standardized
final exam. Finally, Steiner, Holley, Gerdes, and Campbell (2006) found that how much students perceive they learned in a course is an important predictor of SET scores.

However, not all validation measures have proven so successful. Measures of teacher effectiveness by colleagues and administrators, after classroom visitations, are generally unreliable and not systematically correlated with SETs, or other indicators of effective teaching (Centra, 1979; Marsh 1987; Murray, 1980).

Furthermore, it is safe to say that the issue of validity is far from settled. In the following sections the authors discuss some of the major variables that potentially bias student perceptions of faculty teaching.

Reliability

On the question of whether SETs provide reliable measures of teaching effectiveness, there is now a substantial body of evidence: according to Theal and Franklin (2001), “Whether reliability is measured within classes, across classes, over time, or in other ways, student ratings are remarkably consistent” (p. 50).

Marsh (1987) offers extensive evidence of SET reliability both over time, and across various courses, when taught by the same professor. He concludes that “given a sufficient number of students, the reliability of student evaluations compares favorably with that of the best objective tests” (p. 9). Marsh also summarizes the findings of a longitudinal study where a panel of students was asked to evaluate a course at the end of the term, and again several years later. The end of term evaluations of 100 courses correlated .83 with the ratings provided several years later. The median ratings were nearly identical.

Finally, Marsh reports his findings from data collected on more than 1,000 courses including sections of the same course, taught by the same instructor, different courses taught by the same instructor, and the same course taught by different instructors. His analysis showed that student ratings are a significantly better reflection of the effectiveness of the teacher over the influence of the course. When the same instructor taught the same course the correlation between overall instructor ratings was (.72). When the same instructor taught different courses, the correlation was (.61). Finally, when the same course was taught by a different instructor, the correlation was (.05).

Race

There is a small, but growing, number of studies that have examined the effect of instructor race on SET ratings. This issue, while certainly significant in and of itself, may become more important given that the number of African Americans pursuing graduate studies in the United States has risen dramatically (Gabbidon, 2002). Inevitably, this will lead to an increase in the number of African American faculty.

Several recent studies have found that race does matter in SETs. Hendrix (1998) collected data through observation, semi-structured interviews, and open-ended interviews.
with six professors (three black and three white) and 28 students. The purpose of the research was to explore how race affects an instructors’ credibility. When interviewed, students suggested that they apply “more stringent credibility standards to professors depending on a combination of professor race and subject matter” (p.748). Interestingly, Hendrix also found that once African American professors establish credibility, students generally hold favorable attitudes.

Hase (2001) discovered that being a foreign born instructor carried with it both advantages and disadvantages. In recounting her years of teaching American students, Hase describes how her ancestry provided her with instant credibility if the course was in some way related to her identity. However, she also had negative experiences where her identity led to what she perceived as undue criticism. As one example, Hase offers a specific example from one course on global gender issues. In this course, Hase's critical examination of the United States, within the context of the subject matter, resulted in negative experiences that she attributed to being a foreign born instructor.

A more comprehensive study of the effects of race on SETs was conducted by Gabbidon (2002). He compared student evaluations from research methods courses taught at two different institutions. Noting that there was little to no variation in the style and content of the course over the period under study, the major difference between the courses was that some classes were taught at a historically black institution (98% minority students) and the other courses were taught at a primarily white institution. Gabbidon (2002) compared two specific items from the institutions SETs that assessed the overall quality of the instructor. The findings indicated no significant differences between the ratings of the students at the two universities. Gabbidon (2002) also compared the items from the SETs from a race related criminal justice course. Although, significant differences did emerge, he cautions that student interest in the topic, or the subject matter itself, could have produced the differences observed.

The findings of each of these studies on the significance of race in SETs is limited. As Gabbidon (2002) and others acknowledge, their findings are based upon the experiences of only single instructors. Nonetheless, they do address important questions about the influence of race on students’ evaluation of their faculty.

**Gender**

The research focused on the effects of gender in SETs has been mixed. On the one hand, numerous studies have found that female instructors receive lower evaluations than their male peers (Neath 1996; Sandler 1991; Lueck 1993). In fact, Neath (1996), in an article appropriately titled *How to Improve Your Teaching Evaluations Without Improving Your Teachings* suggests that changing one’s gender (if female) will improve one’s effectiveness ratings. However, research conducted by Rowden & Carlson (1996) found that female instructors were rated significantly higher than their male counterparts. These studies demonstrate the lack of consensus about this issue, a point echoed by Marsh and Roche (1997), whose review of research on this topic found inconsistent findings,ulti-
mately leading the authors to conclude that instructor gender has little to no effect on overall student ratings.

**Grades**

One is not likely to find a more controversial issue with regard to the value of SETs than the issue of grades. The oft-heard hypothesis is that inflating grades is the surest way to positive SETs. There is in fact some evidence to suggest that this is true. A number of studies indicate that giving out higher grades translates into higher SETs (Chako, 1983; Neath, 1996; Lersch and Greek, 2001). In fact, the research conducted by Lersch and Greek (2001) suggests a strong association between grades and SET scores. Their study examined course evaluation data from criminology and criminal justice programs at three first tier research universities in the state of Florida. The evaluations of these institutions were the same due to a requirement by the Florida Board of Regents that requires all state universities to use the same evaluation form. In their analysis, grades were the strongest correlate of instructor effectiveness. Although, the researchers go on to note that other significant variables such as the students major, level of interest, perceived fairness of the grading system, and motivation were not included in the study.

Despite these findings, the strength of the relationship between grades and SETs remains in question. A meta-analysis by Ellis (1985) found only low correlations between grades and overall perceived teaching quality. Similarly, Eison (1999), who found a statistically significant correlation between grades and student evaluation reports in 28 out of 50 studies, noted that the mean correlation was relatively weak (.14). Finally, Steiner et al. (2006) note that research consistently suggests that the grade a student believes he/she will receive in a class influences student responses to SETs. Generally speaking, it appears that the overall effect is small, but the nature of the relationship remains poorly understood.

**Class Size and Type of Course**

Researchers have also recognized that variables aside from individual teacher characteristics and grades can influence SETs. In fact, some of these variables have to do with the type and size of the course, variables over which faculty usually have little control. Two such examples that will be discussed here: class size and the type of course being offered.

Research devoted to the effects of class size on SETs has been mixed. While there is some evidence to suggest that smaller classes produce higher SETs, these results are far from conclusive. Gleason (1986) examined 30 studies and found that 20 concluded that smaller class sizes received higher SET scores. He went on to report that even where class size demonstrated a statistically significant effect, the effect was modest.

The type of course (e.g. required or elective) has been a variable of interest in numerous studies. A meta-analysis conducted by Neath (1996) indicated that required courses are bad for SET scores. According to Neath, required courses, “although educationally important, tend not to be well-liked, even by students within the major” (p. 1367). Marsh
and Roche (1997) agree. Their findings suggest that higher evaluations are more common when the reason for taking the course is general interest, or major elective, as opposed to major requirement, or general education requirement.

**Individual Measures of Instructor Performance**

Finally, there are a number of studies that examine individual teacher performance items in relation to overall SET scores. In 1976 Feldman found that five dimensions of teaching had the highest correlations with actual overall evaluation. These dimensions include: stimulation of interest, clarity of explanation, intellectual challenge, sensitivity to class level and progress, and preparation or organization.

More recently Spencer and Schmelkin (2002) found that a group of sophomores, juniors, and seniors attending a private university defined effective teaching as concern for students, valuing student opinions, clarity in communication, and openness toward varied opinions. In 2005, Okpala and Ellis collected data from 218 different colleges in the U.S. regarding student perceptions of teacher quality. Their analysis uncovered five important qualities of effective teaching including caring for students and their learning, content knowledge, verbal skills, dedication to teaching, and teaching skills.

Some researchers have included college faculty themselves in their inquiries into what makes a quality teacher. Schaefer, Epting, Zinn and Buskit (2003) asked 99 faculty, and 231 students, to identify the top 10 qualities of effective college teachers from a list of 28 qualities. Although the orders did vary, both groups agreed upon eight of the top 10 traits: approachable, creative and interesting, encouraging and caring, enthusiastic, flexible and open minded, knowledgeable, realistic expectations and fair, and respectful.

**Statement of Research Problem**

The general goal of this project is to identify the characteristics that make a good college professor. More specifically, our intent is to explore a series of questions drawn from student evaluation forms, and to determine which of the individual items are most strongly tied to overall student perceptions of instructor effectiveness. The items evaluated in this analysis range from general demographic indicators (e.g. student and instructor gender), to class characteristics (e.g. whether the student had a strong desire to take the course, whether the class was in the student’s intended major), to expected course grades, and more specific measures of instructor performance (e.g. whether the instructor was well prepared for class, whether he or she demonstrated concern for student progress).

**Sample and Data**

This research represents a secondary analysis of data collected as part of the teacher evaluation process at a Midwestern university. This university is a Master’s degree granting institution with a total enrollment ranging between approximately 23,000 and 24,000 students over the study period. The School of Criminal Justice (SCJ) grants both Bachelors and Masters degrees in criminal justice, and also features a legal studies program and a
police academy. The annual enrollment of the SCJ was between 700 and 800 students during the period studied. Individual undergraduate classes ranged between a high of 100-150 students for a large section of CJ 101 (justice in society / introduction to criminal justice) to fewer than 20 in research methods classes such as CJ 300, or the senior capstone course (CJ 495).

As part of one of the class meetings held during the final few weeks of the semester, each instructor is expected to distribute course evaluation forms to his or her students. The forms have both a quantitative component (where students respond to series of close ended questions) and a qualitative component (where they can provide open ended comments). While the exact procedure for administration of the evaluations varies from instructor to instructor, students are typically given approximately 15 minutes at the end of a class to complete the survey. The instructor leaves the room during the evaluation period, and the forms are returned to the instructor’s department in a sealed envelope.

For the purpose of this analysis, the authors compiled a sample of all of the student evaluation surveys completed by pupils enrolled in 5 specific courses (CJ 101, 201, 300, 312 and 495 - see Table 1 for course descriptions), taught by instructors in the School of Criminal Justice between the years 2004 and 2009. Over the study period, the SCJ employed approximately 15 full-time, tenured, or tenure track faculty, and a varied number of visiting and adjunct professors. Because the data was identity stripped by both student and instructor, it was not possible for the authors to determine the identity of any faculty. However, in general, over 2/3 of the courses in the SCJ are taught by full time professors: the only class that was likely to have been taught by a part time faculty member on a regular basis would have been selected sections of CJ 101.

A total of 8,117 completed surveys were used to construct the data set for this project. Table 1 presents a description of each course, and the number of surveys drawn from those classes. These five specific courses were chosen because they represent a broad spectrum of criminal justice topics, and because the SCJ teaches multiple sections of each course, to a substantial number of students, every year. This minimized the chances that any individual instructor, or student, could be inadvertently identified as part of this project. Hence, although the sample in question can most accurately be characterized as convenience based, it is likely to be broadly representative of the experiences of criminal justice students, and instructors, at mid-sized Midwestern universities. The size of the sample (n > 8,000) also permits substantial statistical power for quantitative analysis.

**Variables**

**Independent Variables**

In total, the analysis utilized 16 different independent variables, subdivided into 4 conceptual clusters: basic student information (3 variables), basic course/instructor information (4 variables), expected grade in class (1 variable), and student assessments of instructor performance (8 variables). A description of each of these variables is provided below. Basic descriptive data for each indicator can be found in Table 2.
### Table 1. Class Information and Total Enrollment for the Data Set Used in this Study

<table>
<thead>
<tr>
<th>Course Code</th>
<th>Class Description</th>
<th>Total Enrollment Between 2004 and 2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>CJ 101 Justice and Society (Introduction to Criminal Justice)</td>
<td>This introduction to the study of crime and justice includes theories and methodologies from a variety of social science disciplines. The course also provides an introduction to the study of social control and to the origins of crime at the individual, structural and cultural levels.</td>
<td>4988 (61.5% of sample)</td>
</tr>
<tr>
<td>CJ 201 Criminology</td>
<td>An analysis of crime, criminal behavior, punishment, and the theories of deviancy from historical perspectives.</td>
<td>1210 (14.9% of sample)</td>
</tr>
<tr>
<td>CJ 300 Research Methods in Criminal Justice</td>
<td>This course involves an examination of basic investigative methods in criminal justice. Focus is on the logic and theory of criminological research, the formulation and testing of hypotheses, research design, sampling, modes of data production, and the ethics of conducting research in criminology and criminal justice.</td>
<td>523 (6.4% of sample)</td>
</tr>
<tr>
<td>CJ 312 Police Process</td>
<td>Functions of law enforcement and the roles of the police in contemporary society. Study of the police from several perspectives: historical, sociological, psychological, organizational, and political. Issues, research, and trends pertinent to law enforcement organizations.</td>
<td>643 (7.9% of sample)</td>
</tr>
<tr>
<td>CJ 495 Issues in Criminal Justice (Capstone Course)</td>
<td>A capstone course that will entail readings and discussion on contemporary criminal justice issues, ethics, and trends resulting in a senior paper/project.</td>
<td>753 (9.3% of sample)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>n = 8,117</strong></td>
</tr>
</tbody>
</table>

### Basic Student Information

Three student information variables were used as part of this analysis: student gender, student academic status, and whether the class being evaluated was part of the student’s intended major. It was not expected that these variables would have strong influences on perceptions of instructor quality, but it was deemed important to include them as basic controls.

The operationalization of the gender variable was self-explanatory; students were asked to identify if they were male or female. Academic status measured the student’s progress...
### Table 2. Variable Metrics and Bivariate Correlations with the Dependent Variable

<table>
<thead>
<tr>
<th>Variable</th>
<th>Metrics</th>
<th>Non-Parametric Correlation with D.V.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student Gender</td>
<td>4098 female, 3848 male</td>
<td>-.011</td>
</tr>
<tr>
<td>Student Academic Status</td>
<td>3126 freshman, 1783 sophomore, 1426 junior, 1679 senior, 18 graduate</td>
<td>-.065 **</td>
</tr>
<tr>
<td>Class Part of Major</td>
<td>3368 yes, 4604 no</td>
<td>.018</td>
</tr>
<tr>
<td>Instructor Gender</td>
<td>1469 female, 6648 male</td>
<td>.054 **</td>
</tr>
<tr>
<td>CJ 300 Dummy</td>
<td>523 yes, 7594 no</td>
<td>-.085 **</td>
</tr>
<tr>
<td>Strong Desire to Take Course Dummy</td>
<td>5120 yes, 2823 no</td>
<td>.19 ***</td>
</tr>
<tr>
<td>General Education Dummy</td>
<td>4741 yes, 3231 no</td>
<td>.058 **</td>
</tr>
<tr>
<td>Expected Grade in Class</td>
<td>3315 A, 3807 B, 799 C, 45 D, 8 F</td>
<td>.23 ***</td>
</tr>
<tr>
<td>Learned a Great Deal in Class</td>
<td>Mean = 3.36</td>
<td>.60 ***</td>
</tr>
<tr>
<td>Instructor Challenged Thinking</td>
<td>Mean = 3.21</td>
<td>.55 ***</td>
</tr>
<tr>
<td>Instructor Organized</td>
<td>Mean = 3.52</td>
<td>.61 ***</td>
</tr>
<tr>
<td>Instructor Made Clear Presentations</td>
<td>Mean = 3.57</td>
<td>.64 ***</td>
</tr>
<tr>
<td>Clear Objectives</td>
<td>Mean = 3.48</td>
<td>.61 ***</td>
</tr>
<tr>
<td>Responsive to Student Questions</td>
<td>Mean = 3.67</td>
<td>.60 ***</td>
</tr>
<tr>
<td>Concern for Student Progress</td>
<td>Mean = 3.34</td>
<td>.57 ***</td>
</tr>
<tr>
<td>Fair Evaluation Methods</td>
<td>Mean = 3.38</td>
<td>.61 ***</td>
</tr>
<tr>
<td>Overall Instructor Score (D.V.)</td>
<td>Mean = 3.50</td>
<td>-</td>
</tr>
</tbody>
</table>

*** = p < .001, ** = p < .01, * = p < .05

in the program: respondents were asked to self-identify as freshmen, sophomores, juniors, seniors or graduate students (taking undergraduate courses). The intended major variable was measured as a simple dichotomy: “yes” if the class was in the respondent’s intended major, or “no” otherwise.

**Basic Course / Instructor Information**

The second cluster of variables assessed basic course and instructor characteristics. Four individual measures were analyzed:

The first was the gender of the instructor: the operationalization of which is self-evident. The second was a dummy variable (“1” yes, “0” no) assessing whether the class in question was CJ 300 (quantitative research methods). The third was another dummy variable measuring whether the student had a strong desire to take the class, and the last was a dummy variable assessing whether the course was being taken as part of the general education program.
Expected Grade in Class

As previously discussed, it is a popular perception that student evaluation scores are less a valid measure of instructor performance, and more a representation of how much students “like” a particular teacher. To this end, some instructors believe that they can become “liked”, and thus ensure high SET scores, simply by inflating the grades in their classes. To investigate this issue, the present study includes a measure of expected grade as part of the statistical analysis. Students were asked to indicate whether they expected to score an A, B, C, D or F in the class. Unfortunately, the data for this project did not include a measure of the actual grade obtained in the class. However, the authors did review data from their own classes and found that the statistical distributions of expected grades, and their actual grade distributions, were relatively consistent. Where discrepancies existed, they tended to be at the lower end of the scale: that is, few students indicated that they expected to receive D’s or F’s, but a substantial number actually received these grades.

Student Assessments of Instructor Performance

The final eight independent variables are the most theoretically important to this research project. Each of these items represents a dimension of instructor performance. Although there is remarkably little empirical research on the relationships between these types of items and overall perceptions of instructor quality, studies by Feldman (1976), Spencer and Schmelkin (2002), Okpala and Ellis (2005) and Schaefer et al. (2003) do indicate that they may be important. Presumably, instructors who score highly on these 8 items will also be the instructors who obtain the highest overall SET scores.

The individual dimensions of instructor performance are as follows:

First, students were asked whether they learned a lot from the course in question. Second, they were asked whether the instructor successfully challenged their thinking. Then, they were asked to evaluate whether the instructor was organized, whether he or she made presentations that were understandable, and whether or not he or she conveyed clear objectives for the course. Finally, students were asked whether the instructor was responsive to student questions, whether he or she exhibited concern for student progress, and whether the students perceived the methods of evaluation in the class as fair.

For each item, students were asked whether they strongly agreed, agreed, were neutral, disagreed, or strongly disagreed with the statement under consideration. For the purposes of this project, responses were coded on a 5 point scale ranging from “4” (strongly agree) to “0” (strongly disagree). Although the authors recognize that these variables technically represent ordinal Likert scales, for the purpose of the statistical analysis, they were treated as continuous, interval level measures. This permitted simpler presentation and interpretation of the research results and was not likely to substantially violate the technical assumptions of the analytic techniques used (Glass, Peckham and Sanders, 1972).
It should be noted that the individual indicators described here could have been combined to form a single scale with an alpha reliability of .91. However, because the goal of this project was to determine which indicators had the most powerful influences on overall perceptions of instructor quality, this was not done. Instead, each item was entered into the statistical models individually. Because of the large number of cases available for analysis, this did not create serious statistical issues. All of the models could be successfully estimated, and colinearity remained within generally acceptable boundaries. The highest VIFs appeared in the model containing all 16 independent variables and peaked at 2.69: well below the threshold of 5 proposed by most authors (Kutner, Nachtsheim and Neter, 2004).

**Dependent Variable**

The dependent variable in this analysis represented a measure of overall student perception of instructor quality. Students were asked to provide a 5 point response (ranging from “strongly agree” to “strongly disagree”) to the following question: The quality of the instructor for this course was excellent (try to set aside your feelings for the course). Once again, although this variable technically represents an ordinal level measure, it was treated as an interval scale measure for the purpose of the statistical analyses.

**Research Hypothesis**

The following section outlines the specific sets of research hypotheses that were tested in this study. The goal of each model is to determine whether the variables analyzed during that step were significantly associated with overall perceptions of instructor effectiveness. The last model, which contains all of the independent variables, provides information for assessing the relative importance of the indicators under consideration.

**Model One: Basic Student Information**

This regression model tests three null hypotheses. They are as follows: controlling for the other variables in the model, the gender of the student, his or her academic status, and whether the course under consideration is in the student’s intended major, will have statistically non-significant effects on that student’s perception of the instructor’s effectiveness.

**Model Two: Basic Course / Instructor Information**

This regression model is similar to the first. It tests the following four null hypotheses: controlling for the other variables in the model, the instructor’s gender, the student’s desire to take the class, his or her taking it to fulfill general education requirements, and whether the course under evaluation is CJ 300 will have statistically non-significant effects on that student’s overall perception of the class instructor’s effectiveness.
Model Three: Basic Student, Course and Instructor Characteristics + Expected Grade in Course

Model three incorporates all of the variables from models one and two. In addition, this model tests the null hypothesis that controlling for the seven other items in the analysis, the student’s perceived grade in the class will have no significant effect on that student’s rating of the effectiveness of the course instructor.

Model Four: Basic Student, Course and Instructor Characteristics, Expected Grade, and Individual Indicators of Instructor Performance

Model four incorporates all of the variables from model three. In addition, it tests a series of research hypotheses pertaining to eight individual measures of instructor performance:

1) To the extent that a student perceives that he or she “learned a great deal” from a class, that student’s rating of the overall effectiveness of the instructor will be higher.
2) To the extent that a student perceives that the class “challenged his or her thinking”, that student’s rating of the overall effectiveness of the instructor will be higher.
3) To the extent that a student perceives that the instructor’s methods of evaluation for the class were fair, that student’s rating of the overall effectiveness of the instructor will be higher.
4) To the extent that a student perceives that the instructor exhibited concern for student progress, that student’s rating of the overall effectiveness of the instructor will be higher.
5) To the extent that a student perceives that the instructor was responsive to student questions, that student’s rating of the overall effectiveness of the instructor will be higher.
6) To the extent that a student perceives that the instructor made understandable presentations in the class, that student’s rating of the overall effectiveness of the instructor will be higher.
7) To the extent that a student perceives that the instructor ran well organized class sessions, that student’s rating of the overall effectiveness of the instructor will be higher.
8) To the extent that a student perceives that the instructor clearly communicated course objectives, that student’s rating of the overall effectiveness of the instructor will be higher.

Given the logic of multivariate regression analysis, each of the following effects is expected to hold, controlling for all of the other variables in the model.

Analysis

The models were tested using Multivariate, Ordinary Least Squares (OLS) regression analysis (also known as “linear regression”). OLS was chosen for this study because it
provides intuitively appealing, and easily interpretable, statistical estimators. This is not always true of other, more complex types of regression techniques. Although the dependent variable in this case is an ordinal scale (as opposed to the interval / ratio level measure typically analyzed using OLS), it was deemed to be a reasonable approximation to a normally distributed measure. This decision is supported by the fact that OLS is widely known to be “robust for validity against nonnormality” (PROPHET Statguide, 1997). Similarly, although mathematically transforming the dependant variable could have normalized its statistical distribution (PROPHET Statguide, 1997), it would have made it impossible to interpret the effects of the independent variables on instructor quality in their original metrics. This would have made the results far less accessible to a general audience, and consequently, the present authors elected not to undertake such transformations.

**Results**

Table 3, presented below, reveals the results of testing the basic student information hypotheses described above:

<table>
<thead>
<tr>
<th>Variables</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>3.59</td>
<td>.021</td>
<td>170.57</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>What is your gender? (Male)</td>
<td>.005</td>
<td>.018</td>
<td>.28</td>
<td>.78</td>
</tr>
<tr>
<td>What is your academic status?</td>
<td>-.042</td>
<td>-.008</td>
<td>-4.98</td>
<td>.000</td>
</tr>
<tr>
<td>Is this course in your intended major?</td>
<td>.000</td>
<td>.020</td>
<td>-.009</td>
<td>.99</td>
</tr>
</tbody>
</table>

\[ r^2 = .004 \]

The statistics in this table indicate that the gender of the student, and whether or not the course under evaluation is in the student’s major, are not significantly associated with perceptions of overall instructor effectiveness. In other words, there was insufficient evidence to reject the null hypotheses. However, academic status is significantly linked to instructor effectiveness \((p < .001)\). It seems that more senior students have higher expectations than their junior counterparts. For each step a student advances up the seniority ladder (i.e. from freshman to sophomore, sophomore to junior, junior to senior, senior to graduate student) his or her evaluation of the instructor declines by .04 of a point. Conceptually, this does not represent a huge difference: it suggests that, on average, seniors are approximately .12 of a point “harsher” than freshman. However, the difference has strong statistical significance given the large size of the sample (n = 7,796 for this analysis). Table 4 reveals the results of testing the basic course/instructor related hypotheses described above.
Table 4. The Effects of Basic Course and Instructor Characteristics on Overall Instructor Effectiveness

<table>
<thead>
<tr>
<th>Variables</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>3.14</td>
<td>.027</td>
<td>114.73</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Did you have a strong desire to take this course?</td>
<td>.32</td>
<td>.019</td>
<td>.19</td>
<td>16.74</td>
</tr>
<tr>
<td>Methods Class (CJ 300)</td>
<td>-.025</td>
<td>.038</td>
<td>-.008</td>
<td>-.65</td>
</tr>
<tr>
<td>Are you taking this course to fulfill general education requirements?</td>
<td>.14</td>
<td>.019</td>
<td>.083</td>
<td>7.23</td>
</tr>
<tr>
<td>Gender of Instructor (Male)</td>
<td>.091</td>
<td>.023</td>
<td>.044</td>
<td>3.91</td>
</tr>
</tbody>
</table>

r-squared = .044

| F=89.29 (DF = 4) | Sig. < .001 |

The results here indicate the following: Students enrolled in CJ 300 (quantitative research methods) do not appear to rate their instructors any more harshly than students enrolled in other courses. However, the other three measures are significantly related to overall perceptions of instructor effectiveness ($p < .001$). Students who have a strong desire to take a particular class, those who are taking classes to fulfill general education requirements, and those who have male instructors, are all more likely to rate their instructors positively. Of these effects, the “strong desire to take the class” has the most predictive power (as evidenced by the standardized Beta coefficients). Students who indicated that they had a strong desire to take a class gave their instructors mean evaluation scores that were over .3 of a point higher than students who indicated that they did not have strong desire to take that class.

Table 5 presents the results of the regression model that combined all of the predictors from Tables 3 and 4, and also added a variable measuring the student’s perceived grade in the class being evaluated.

This analysis reveals the following: the gender of the student, his or her academic status, whether the class is in the student’s intended major, and whether the class under evaluation is CJ 300 have no significant impact on the overall instructor effectiveness score (once all of the other variables in the model have been held constant). The reader will recall that these results are consistent with the analyses presented in Tables 3 and 4, except that in Table 3 academic status was a significant predictor of perceived instructor effectiveness. Interpretations of this (and other interesting findings) will be presented in the discussion section. Having a strong desire to take the class, taking the class with a male instructor, and taking the class to fulfill general education requirements continue to have significant, and positive, associations with overall instructor effectiveness ($p < .001$ in all cases). These findings are consistent with the results presented earlier.
Table 5. The Effects of Basic Student, Course and Instructor Characteristics, and Perceived Grade on Overall Instructor Effectiveness

<table>
<thead>
<tr>
<th>Variables</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>2.36</td>
<td>.058</td>
<td>41.04</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>What is your gender?</td>
<td>.018</td>
<td>.018</td>
<td>.011</td>
<td>1.01</td>
</tr>
<tr>
<td>What is your academic status?</td>
<td>-.007</td>
<td>.009</td>
<td>-.011</td>
<td>-.83</td>
</tr>
<tr>
<td>Is this course in your intended major?</td>
<td>-.011</td>
<td>.022</td>
<td>-.007</td>
<td>-.49</td>
</tr>
<tr>
<td>Did you have a strong desire to take this course?</td>
<td>.29</td>
<td>.020</td>
<td>.17</td>
<td>14.55</td>
</tr>
<tr>
<td>Methods Class (CJ 300)</td>
<td>.05</td>
<td>.039</td>
<td>.016</td>
<td>1.29</td>
</tr>
<tr>
<td>Are you taking this course to fulfill general education requirements?</td>
<td>.14</td>
<td>.023</td>
<td>.087</td>
<td>6.19</td>
</tr>
<tr>
<td>Gender of Instructor</td>
<td>.08</td>
<td>.023</td>
<td>.039</td>
<td>3.54</td>
</tr>
<tr>
<td>Expected Grade</td>
<td>.25</td>
<td>.013</td>
<td>.21</td>
<td>18.60</td>
</tr>
</tbody>
</table>

$r^2$-squared = .086  
F=90.90  
(DF = 8)  
Sig. < .001

Finally, and perhaps most importantly, the present analysis shows that students who expect to obtain higher grades in a class rate their instructors significantly more generously! For each one grade level change (i.e. from an “F” to a “D” from a “D” to a “C” and so forth) the overall instructor effectiveness score improves by approximately 0.25 of a point. Moreover, the standardized Beta coefficient for this variable (.21) is larger than that for any other variable under consideration. This represents *prima facie* evidence that “easier” instructors are better liked, and tend to obtain higher scores on student evaluations (all else being equal). The differences are conceptually important, students who expect to get an “F” in a class are likely to rate their instructors a full point lower than students who expect to obtain an “A”. However, it is important to note that direct measures of instructor performance have not yet been included in this model. Table 6 presents the results of the regression analysis which did explore this issue.

In the composite model containing all 16 independent variables, the following measures are *not* statistically associated with overall perception of instructor effectiveness: academic status, intended major, strong desire to take the course, the dummy variable for CJ 300 and, most notably, expected grade in course. The fact that one's anticipated grade in the course is no longer significant is a statistically, and conceptually, important difference from Table 5. After the individual instructor performance factors are held constant, students who expect to get high grades in classes no longer rate their teachers any higher than those who expect to receive low grades.

Consistent with the previous models, students in this analysis continue to rate male instructors slightly, but still significantly, higher than female instructors ($p = .01$) and
Table 6. The Effects of Basic Student, Course and Instructor Characteristics, Perceived Grade, and Individual Instructor Performance, on Overall Instructor Effectiveness

<table>
<thead>
<tr>
<th>Variables</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>(Constant)</td>
<td>-.28</td>
<td>.047</td>
<td></td>
<td>-6.03</td>
</tr>
<tr>
<td>What is your gender?</td>
<td>.02</td>
<td>.011</td>
<td>.014</td>
<td>1.97</td>
</tr>
<tr>
<td>What is your academic status?</td>
<td>.001</td>
<td>.006</td>
<td>.002</td>
<td>.21</td>
</tr>
<tr>
<td>Is this course in your intended major?</td>
<td>-.009</td>
<td>.014</td>
<td>-.005</td>
<td>-.62</td>
</tr>
<tr>
<td>Did you have a strong desire to take this course?</td>
<td>-.02</td>
<td>.013</td>
<td>-.012</td>
<td>-1.54</td>
</tr>
<tr>
<td>Methods Class (CJ 300)</td>
<td>.04</td>
<td>.024</td>
<td>.012</td>
<td>1.64</td>
</tr>
<tr>
<td>Are you taking this course to fulfill general education requirements?</td>
<td>.034</td>
<td>.014</td>
<td>.021</td>
<td>2.38</td>
</tr>
<tr>
<td>Gender of Instructor</td>
<td>.036</td>
<td>.014</td>
<td>.017</td>
<td>2.48</td>
</tr>
<tr>
<td>Expected Grade</td>
<td>-.013</td>
<td>.009</td>
<td>-.011</td>
<td>-1.47</td>
</tr>
<tr>
<td>Learned Great Deal in Class</td>
<td>.20</td>
<td>.01</td>
<td>.20</td>
<td>19.24</td>
</tr>
<tr>
<td>Material Challenged Thinking</td>
<td>.11</td>
<td>.009</td>
<td>.12</td>
<td>12.94</td>
</tr>
<tr>
<td>Methods of Evaluation Fair</td>
<td>.14</td>
<td>.01</td>
<td>.14</td>
<td>13.94</td>
</tr>
<tr>
<td>Instructor Concerned Student Progress</td>
<td>.11</td>
<td>.009</td>
<td>.11</td>
<td>11.65</td>
</tr>
<tr>
<td>Instructor Responsive to Questions</td>
<td>.17</td>
<td>.013</td>
<td>.13</td>
<td>13.13</td>
</tr>
<tr>
<td>Instructor Presentation Understandable</td>
<td>.20</td>
<td>.013</td>
<td>.17</td>
<td>14.90</td>
</tr>
<tr>
<td>Instructor Organized Class Sessions</td>
<td>.10</td>
<td>.013</td>
<td>.09</td>
<td>7.69</td>
</tr>
<tr>
<td>Course Objectives Clearly Communicated</td>
<td>.08</td>
<td>.012</td>
<td>.07</td>
<td>6.32</td>
</tr>
</tbody>
</table>

r-squared = .65
F=859.15 (DF = 16) Sig. < .001

classes taken as part of the general education program somewhat higher than those not taken to satisfy general education requirements (p=.02). However, contrary to previous models, the sex of the student now significantly impacts perceptions of instructor performance (with males giving marginally higher scores, p = .05), while having a strong desire to take a course is no longer a statistically significant predictor (the reader should recall that this was the strongest significant predictor in Table 4).

The most striking findings from the present analysis; however, pertain to the eight individual instructor performance variables. All of these variables are strongly, and positively associated with the overall instructor evaluation score (even after controlling for the other variables in the model, p < .001 in all cases). Students who agreed that they learned a great deal in class, that the class challenged their thinking, that the instructor’s methods of evaluation were fair, that the instructor demonstrated concern for their progress, that he or she was responsive to student questions, that he or she made understandable presentations, that class sessions were well organized, and that course objectives

©2011 All rights reserved
were clearly communicated, consistently rated their instructors higher than those who disagreed with these statements. The two strongest items (as measured by the Beta coefficients) turned out to be “learned a great deal in class” and “instructor made understandable presentations”. For each extra point on these two items (i.e. moving from “agree” to “strongly agree” and so forth) students assigned their instructors approximately .2 of an extra point on their overall evaluations.

Discussion

The most important findings from the regression analysis are that although student/course/instructor characteristics like gender, nature of the class under consideration, etc. can influence overall perceptions of instructor effectiveness, their overall impact is small: particularly when compared to more direct measures of instructor performance. Students who expect to score high grades in classes are more likely to rate their instructors higher in the statistical model that controls only for basic characteristics (Table 5); however, they are not any more likely to favor their instructors after measures of instructor performance have been held constant (Table 6). This suggests that what an instructor chooses to do in class is more important to his or her overall SET scores than more superficial variables. More specifically, this study provides statistical evidence to support the idea that well prepared, organized, caring, responsive, fair instructors will be rated higher than those who struggle in these areas. These finding are consistent with the research conducted by Crumbley, Henry, and Kratchman (2001) and helps to validate subjective instructor evaluations as a legitimate measure of teaching effectiveness (See also Spencer & Schmelkin, 2002; Okpala & Ellis, 2005; and Schaefer et al., 2003). Our study also helps dispel myths like: teachers can “buy” high evaluation scores with high grades, or teachers who are assigned “difficult” classes like quantitative research methods are doomed to receive low evaluation scores.

Given the variation between these findings and others, research in this area should seek to establish a reliable estimate of the strength of the relationship between grades and overall SETs. It is possible that high grades are actually a result of effective teaching, as, students who learn more tend to give their teachers higher overall evaluations (Marsh & Roche 1997). However, to the extent grades significantly effect student perceptions of effective teaching, questions about the value of SETs will remain.

The finding that male instructors seem to be rated consistently higher than female instructors is also conceptually interesting. Perhaps it reflects some inherent bias against women teaching at post secondary institutions. However, the present researchers would like to offer a word of caution as these findings are based on evaluations of faculty in a school of criminal justice. As such, these findings may be less applicable in other disciplines given that criminal justice is traditionally a male dominated discipline. It may be possible that students do have a gender bias when the subject is criminal justice: much like Gabbidon (2002) and Hase (2001) found with classes pertaining to race and culture taught by minority professors.

Based on the present analysis, it is possible to assert that this gender effect does not reflect that male instructors simply perform better in the classroom than females. Were this...
true, one would have expected this effect to statistically “disappear” once direct measures of instructor performance were controlled in Table 6: yet this did not occur. Unfortunately, this study does not provide any additional data to examine possible explanations for this phenomenon. Moreover, because our sample was drawn from only one Midwestern university, there is no guarantee that even this basic finding will generalize to other institutions. As noted earlier, the literature has provided inconsistent results in this area. Consequently, it seems most appropriate at present to merely note this finding as interesting.

Conclusions and Limitations of the Present Study

Overall, this study provides preliminary evidence to support the notion that SET evaluations may be a valid, and reliable, method of assessing instructor performance. The analysis suggests that faculty who care about their students, are well organized, perceived as fair, and who inspire student thinking, seem to be perceived as the most effective classroom teachers. These instructor performance items cluster quite reliably (alpha = .91), and are more strongly connected to overall perceptions of teacher quality than descriptive student, course, or instructor characteristics. They also seem to matter more than the ultimate grades that students expect to receive in their classes.

When taken together, these findings provide an encouraging message: they suggest that students want, and expect, far more out of their classes than simply “getting an A”. They want instructors who value them, feel passionate about the subjects that they teach, and are able to inspire student learning. As college tuition continues to get more expensive, and employers continue to become more demanding of college graduates, it seems plausible that these trends will continue. When one is paying thousands of dollars for an education, it seems only reasonable that one would expect to walk away with more than “a piece of paper” provided by an indifferent faculty and/or institution.

That said, the authors do not wish to claim too much for the present analysis. The speculative conclusions offered in the previous paragraph assume that one takes the findings at face value. There are a number of methodological issues with the present research that warrant caution (particularly when applying the results to other post-secondary institutions).

First and foremost, the reader should recall that this study is based on a non-probability sample of criminal justice students enrolled in a single SCJ, at a single Midwestern university. To the extent that other institutions, and their student populations, are substantially different, these findings may not generalize. More specifically, the present authors would recommend caution when applying the findings to prominent tier one research universities, community colleges, or small liberal arts colleges (i.e. places that are substantially different from the university under study). Administrators at institutions where most classes are taught by part time faculty (i.e. doctoral students, adjunds etc.) should also be cautious about assuming that these findings will apply to their schools, as should those where much of the student body consists of foreign, or minority students (as this
university draws primarily from one region in one state, and has a student body that is disproportionately White and middle class.

It should also be noted that the present findings are cross-sectional. Therefore, readers should be cautious in assuming that any of the associations revealed here represent true, causal, relationships. For example, it is the present author’s assumption that the “grade effect” revealed in Table 5 is spurious to the individual instructor performance items analyzed in Table 6. That is, that instructors who grade fairly, inspire their students, make interesting presentations etc. will inspire their students to earn higher grades. However, it is also plausible that students who have earned high grades will be retrospectively more likely to rate their instructors as fair, inspiring, interesting etc. Because the variables in this study are measured at one point in time, the present researchers have no way of determining which causal sequence is more accurate.

Finally, it is clear that some of the measures used in this study are not ideal. Most notably, to be truly confident about “the grade effect” it would have been useful to have a measure of actual grades earned in the class, as opposed to perceived grades. As noted earlier, the authors’ informal analysis suggested there is a fair bit of overlap in these measures within the authors’ own classes; however, this does not necessarily have to be the case with all instructors. It is not implausible that some instructors give students the impression that they will receive grades that are substantially lower (or higher) than the actual grades that appear on their transcripts. Further validating this, and the other measures in this paper, would allow researchers to have a higher degree of confidence in the findings of this research.

Despite these limitations, it does not appear that any of these shortcomings are sufficiently serious as to render the results of this paper meaningless. On the contrary, the present authors would like to suggest that this study makes an important contribution to the debate on the validity of SETs as measures of instructor performance. However, the researchers suggest further inquiry in this area to shed more light on the issue of how student evaluations can best be used as formative, developmental, and summative evaluation tools, within the discipline of criminal justice.

References


Ontario, Canada: Ontario Confederation of University Faculty Associations.
Teaching. Psychological Reports. 78:1363-1372
Okpala, C. O., & Ellis, R. (2005) The perceptions of college students on teacher quality: 
A focus on teacher qualifications. Education, 126, 374-378.
PROHPHET Statguide. (1997). Online resource provided by Northwestern University, 
Evanston: IL.
Rowden, G. and Carlson, R. (1996). “Gender Issues and Students’ Perceptions of In- 
structors’ Immediacy and Evaluation of Teaching and Course.” Psychological Re- 
ports 78: 835-839.
Higher Education 12, 4, 395-401.
Sandler, B. R. (1991). Women Faculty at Work in the Classroom, or Why it Still Hurts 
to be a Woman in Labor. Communication Education. 40: 6-15.
Schaeffer, G., Epting, K., Zinn, T., & Buskit W. (2003). Student and faculty perceptions 
tin, Oct. 1993, pp. 6-8, 12.
Listening to Students While Acknowledging Bias. Journal of Social Work Educa- 
tion, 42(2), 355-376.
Theal, M., & Franklin, J. (2001). Looking for Bias in All the Wrong Places: A Search for 
Truth or a Witch Hunt in Student Ratings of Instruction. New Directions for Institu- 
tional Research, 109, 45-56.
Learning by Doing: An Empirical Study of Active Teaching Techniques

Jana Hackathorn\textsuperscript{a}, Erin D. Solomon\textsuperscript{b}, Kate L. Blankmeyer\textsuperscript{b}, Rachel E. Tennial\textsuperscript{b}, and Amy M. Garczynski\textsuperscript{b}
\textsuperscript{a} Murray State University, Murray, KY 42071
\textsuperscript{b} Saint Louis University, St. Louis, MO 63103

Abstract

The current study sought to examine the effectiveness of four teaching techniques (lecture, demonstrations, discussions, and in-class activities) in the classroom. As each technique offers different benefits to the instructor and students, each technique was expected to aid in a different depth of learning. The current findings indicated that each teaching technique has its own unique benefits and is effective for various levels of learning. Additionally, our findings supported the notion that active techniques do aid in increasing learning. In-class activities led to higher overall scores than any other teaching method while lecture methods led to the lowest overall scores of any of the teaching methods. The implications for the classroom are discussed.

Keywords: Active learning, Bloom’s taxonomy, assessment, teaching techniques.

Traditionally, college lectures consist of teachers verbally communicating information to the students, and students passively receiving and encoding it in their memories (Boyer, 1990; Michel, Cater III, & Varela, 2009; Stewart-Wingfield & Black, 2005). In a typical college classroom, this presents as a teacher lecturing at the front of the room while students feverishly take notes. However, it is probably more likely that most instructors do not solely teach in this passive fashion but also have engaging or interactive classroom moments or situations. Perhaps this is because many recent studies (e.g. Bonwell & Eison, 1991; Michel, et al., 2009) suggest that the passive method may not be the most effective way for students to learn. Rather, current research advocates for teaching techniques that encourage students to actively engage in the material because classroom engagement has been found to promote deeper levels of thinking and better facilitate encoding, storage, and retrieval than traditional lecture (McGlynn, 2005; Peck, Ali, Matchock, & Levine, 2006). Consequently, it is likely that most instructors attempt to incorporate techniques that involve the students and get students thinking about and applying the material (see Michel, et al., 2009 for a review). These techniques can range from demonstrations, to discussions, to in-class activities. Simply put, traditional ideas of lecture have developed a bad reputation, and some may be ready to banish them from their teaching repertoire.

\textsuperscript{1} Corresponding author's email: jhackathorn@murraystate.edu
Active Teaching

Active, or experiential, teaching is a student-centered approach to teaching. It includes any technique that involves the students in the learning process and holds students responsible for their own learning (Bonwell & Eison, 1991; Michel, et al, 2009; Yoder & Hochevar, 2005). Instructors may have a vast arsenal of active teaching techniques at their disposal, perhaps without even being aware of them (e.g. asking questions as part of one’s normal lecture style). Instructors have used elaborate demonstrations, structured activities, journaling, small group discussions, quizzes, interactive lecture cues, videos, humorous stories, taking field trips, and games, to get students involved and active in the learning process (Bonwell & Eison, 1991; Cook & Hazelwood, 2002; Ebert-May, Brewer, & Allred, 1997; Hackathorn, et al., 2010; Michel et al., 2009; Peck, et al., 2006; Sarason & Banbury, 2004).

While the literature on teaching effectiveness is vast, a large portion of the literature has been focused on the effectiveness, or perceived effectiveness, of interactive teaching strategies. These strategies can range from appropriate use of media and electronic resources (Serva & Fuller, 2004) to homework assignments (Bolin, Khramtsova, & Saarnio, 2005) and quizzes (Crone, 2001) to demonstrations (Zaitsev, 2010) and group projects (Kreiner, 2009). For example, Hackathorn and colleagues (2010) used interactive lecture cues, such as prompting students to link the material to personal stories, and found that it was an effective way of increasing students’ depth of learning. Forrest (2005) took her students on a field trip to a hockey game, allowing them to see psychological principles, such as conformity and in-group bias, firsthand. Other instructors have created in-class games based on television game shows like “Jeopardy” (Binek-Rivera & Mathews, 2004) and “Who Wants to Be a Millionaire?” (Cook & Hazelwood, 2002; Saranson & Banbury, 2004) to increase student involvement and enthusiasm in the classroom.

From an innovation point of view, active teaching techniques change the pace of the classroom, and are a creative way to increase students’ involvement, motivation, excitement, attention, and perceived helpfulness and applicability of the class (Binek-Rivera & Mathews, 2004; Bonwell & Eison, 1991; Guthrie & Cox, 2001; Stewart-Wingfield & Black, 2005). From a cognitive perspective, experientially taught students may engage in higher-order thinking such as analysis, synthesis, and evaluation (Anderson & Krathwohl, 2001; Bloom, Engelhart, Furst, Hill & Krathwohl, 1956; Bonwell & Eison, 1991; Hackathorn, et al., 2010). They are also better able to identify the concepts in the real world, manipulate phenomena for their own purposes, think about the material in new and complex ways, comprehend phenomena conceptually, and recall, retain, and memorize the material better (Donovan, Bransford, & Pellegrino, 1999; Driscoll, 2002; Rubin & Hebert, 1998; Serva & Fuller, 2004; Whetten & Clark, 1996).

Although it seems that active teaching strategies should be adopted in every classroom, the literature is still mixed on its effectiveness (see Michel, et al., 2009 for a review). This may be because the majority of the early research studying the effectiveness of teaching techniques are either qualitative in nature (Berger, 2002), anecdotal (Forrest, 2009), used satisfaction or course evaluations (Serva & Fuller, 2004), or used student...
completed, self-report measures of perceived learning (Benek-Rivera & Matthews, 2004) instead of actual cognitive outcomes. While it is important to understand how the students perceive and appreciate active teaching, a cognitive outcome offers a concrete evaluation of the degree to which students have learned a given concept (Tomcho & Foels, 2008).

Bloom’s cognitive processing taxonomy is a valid, reliable, efficient, and effective means of evaluating learning (Anderson & Krathwohl, 2001; Bloom, et al., 1956; Lord & Baviskar, 2007; Noble, 2004). Specifically, the first three levels of Bloom’s taxonomy (knowledge, comprehension, and application) can be used to effectively assess cognitive outcomes, because each level assesses learning at a different depth. The most basic level (i.e. knowledge) mostly assesses the students’ abilities to remember material through questions that prompt students to identify, list, or describe a concept. Second level (i.e. comprehension) items prompt students to reword information in a meaningful manner to show that they understand the material. Third level (i.e. application) items instruct students to apply the material to new phenomena or constructs, which demonstrates their ability to select appropriate information from situations (Anderson & Krathwohl, 2001; Bloom et al., 1956; Granello, 2001; Lord & Baviskar, 2007).

In the past decade, a large number of studies have begun to empirically examine the cognitive effects of active teaching techniques on learning outcomes (e.g. Benek-Rivera & Matthews, 2004; Cook & Hazelwood, 2002; Ebert-May et al., 1997; Sarason & Banbury, 2004; Seipel & Tunnell, 1995; Strow & Strow, 2006; Tomcho & Foels, 2008). However, the results are mixed and often contradictory (see Michel, et al., 2009 for a review). For example, some empirical studies demonstrate that active teaching techniques are superior to lecture (Serva & Fuller, 2004; Michel, et al., 2009; Van Eynde & Spencer, 1988), while others suggest that there is no real difference (Dorestani, 2005; Miner, Das, & Gale, 1984; Stewart-Wingfield & Black, 2005). Thus, further research is warranted.

Perhaps one reason for such mixed results is that many of the empirical studies treat one class of students as an active teaching class (“active”) and compare it to another class of students that emphasizes lectures (“passive”), with the two courses commonly being taught by two separate instructors (Michel, et al., 2009). While overall, this provides evidence either in favor of or against active teaching, it confounds the comparison of the effectiveness of the technique itself. For example, Michel and colleagues (2009) found students in the “active” course were better at learning and memorizing course material than students in the “passive” course. However, because the class and instructors were different, a direct comparison of active teaching and traditional lecture is difficult. The differences may be due to the teaching techniques, the students who self-selected the course or the instructor, the instructor, or some other difference between the groups. Additionally, the authors used a large variety of techniques, without clear operational definitions of where one technique ends and another begins. Michel and colleagues (2009) described their ‘active’ class as containing quizzes, critical thinking exercises, demonstrations, discussions, and in-class activities. However, it is unclear which particular technique was the most effective, or whether one technique accounted for the difference in the learning outcomes. In another example, Stewart, Myers, and Culley (2010) examined
the effectiveness of active teaching through a specific technique of in-class writing assignments. However, the authors noted that in conjunction with the in-class writing assignments discussion was often used. Thus, there is no way to truly discern which was the effective technique, the writing assignments or the discussion.

The Current Study

In order to add to the literature on the effectiveness of active teaching techniques, the current study empirically examined several commonly used active teaching techniques. The current study used the same classroom and instructor to compare various techniques, while also distinguishing between techniques. Four separate teaching techniques (i.e. lectures, demonstrations, discussion, and in-class activities) were used to teach various constructs throughout an entire semester of a social psychology course.

Lectures. Lecturing, sometimes referred to as the “information dump” is a commonly used approach that involves presenting specific information for the majority of class time, allowing little opportunity for student interaction and expects students to have mastered the information by the time of the exam (Stewart-Wingfield & Black, 2005; Whetten & Clark, 1996). Generally, lectures consist of instructors introducing constructs and their definitions, examples of how phenomena work, and other supporting information. This approach is beneficial because it is a convenient and efficient way to introduce a vast amount of information, especially in large classes where activities may be impractical (Michel et al., 2009; Miner et al., 1984; Whetten & Clark, 1996; Van Eynde & Spencer, 1988). Consequently, lecturing has developed a reputation of being mundane, disengaging, or monotonous (Dorestani, 2005; Miner et al., 1984; Stewart-Wingfield & Black, 2005). Some instructors worry that students retain less of the information, and many instructors find themselves dealing with students who pay less attention, play games or send messages on their laptops, or even sleep in class (Michel et al., 2009; Van Eynde & Spencer, 1988).

Demonstrations. Demonstrations involve activities that occur in the classroom as a means of demonstrating how a phenomena ‘works’ (Dunn, 2008). This technique is slightly more active than lecture because the students are able to get involved and see first-hand how the construct or phenomena presents itself in the real world. Additionally, demonstrations can break up the pace of the classroom while also providing an enjoyable experience for the students (Forsyth, 2003). However, generally, demonstrations only engage a few of the students in the classroom, have guidelines and parameters dictating the path of the learning process, and usually lead to a very specific, often predetermined, outcome. For example, in one demonstration, three students are asked to come to the front of the room and identify the flavors of jellybeans to demonstrate the domination of the olfactory bulb on taste. As part of the demonstration, one student is instructed to eat a jellybean normally, one student is instructed to shut his or her eyes while eating the jellybean and the third student is instructed to shut his or her eyes while also plugging his or her nose while eating the jellybean. As the third person is often unable to identify even the strongest flavored jellybeans, this demonstration is an excellent, usually infallible, and sometimes humorous way to illustrate the importance that smell has on our ability to
taste. However, this demonstration does not allow all students to experience the phenomena. Thus, the uninvolved students are still just passively receiving information.

**Discussion.** Discussion, a hybrid form of teaching because students give and receive information, is often considered the prototypic method and core component of active teaching and learning (McKeachie, 2002; Stewart, et al., 2010; Whetten & Clark, 1996). A classroom discussion is an active teaching technique because it enables students to explore issues of interest, opinions, and ideas. However, it also leads to deeper levels of learning because in order to build on each other’s ideas, the students must first listen and understand the contributions of others students in order to respond or add to it (Hadjioannou, 2007). Additionally, past studies have shown that during discussion students are attentive, active, more engaged, and motivated (see Bligh 2000 for a review; Ryan & Patrick, 2001).

**In-class activities.** Arguably, the most active teaching technique is the in-class activity (Whetten & Clark, 1996). In-class activities are usually a technique that involves all of the students in the class, either working in groups or alone, to solve a problem or puzzle. The benefit of an in-class activity is the same as demonstrations, in that it increases attention and students are able to see a phenomena unfold, but are also able to personally manipulate and practice using that phenomena in a first-hand environment (Forsyth, 2003). This is advantageous because students may not truly understand a concept until they have manipulated it for themselves (Whetten & Clark, 1996). Examples of in-class activities can range from playing games as exam reviews (Cook & Hazelwood, 2002; Saranson & Banbury, 2004) to in-class journaling (Bolin, Khramtsova, & Saarnio, 2005).

In the current study, student learning was assessed by administering quizzes and exams that assessed concepts on three levels of Bloom’s taxonomy (i.e. knowledge, comprehension, and application). This methodology allowed the researchers to examine the effectiveness of each individual technique on three depths of learning while also examining the overall effectiveness of the techniques in a comparative fashion. There were five main expectations for the current study.

**Hypothesis 1.** As lecture (LECT) is considered the least effective in helping students learn material (Michel, et al., 2009; Van Eynde & Spencer, 1988), it was expected that for constructs taught using lecture, students might be able to retain or recognize vocabulary words, but may not understand the intricacies or applicability for most phenomena. Thus, it was hypothesized that students would score a higher percentage of correct answers on knowledge level questions than comprehension or application, when constructs were taught using LECT.

**Hypothesis 2.** Although there is evidence that demonstration (DEMOs) increase attention and enjoyment (Forsyth, 2003) as they only allow for minimal interaction as they often only employ a few students from the classroom, it was expected that students may be able to understand the concepts, but may not necessarily have increased memory for vocabulary or an increased ability to apply the concept for themselves. Thus, it was hypothe-
sized that for constructs taught using DEMO, students would score a higher percentage of correct answers on comprehension level questions than knowledge or application.

**Hypothesis 3.** As discussion (DISC) has the potential to involve all students in the activity and that students understand what has been said, in order to contribute (Hadjioannou, 2007), a logical inference is that discussion is probably more effective for comprehension level learning. Thus, it was hypothesized that for constructs taught using DISC students would score a higher percentage of correct answers on comprehension level questions, than knowledge or application.

**Hypothesis 4.** An in-class activity (ICA) allows each student to actually manipulate and practice applying the information for his or her self (Forsyth, 2003; Whetten & Clark, 1996). However, in order to correctly apply the information, one must also understand the material (Bloom, et al., 1956). Thus, it was hypothesized that for constructs taught using ICA, students would score a higher percentage of correct answers on both comprehension and application level questions than knowledge questions.

**Hypothesis 5.** Based on past studies and arguments (McGlynn, 2005; Peck, Ali, Matchock, & Levine, 2006), the authors expected that as the technique became increasingly active, so would the scores on test items. Thus, it was expected that students’ overall scores would be significantly higher for constructs taught using ICA than LECT methods.

**Method**

During a social psychology course, various constructs were taught using one of the aforementioned techniques: LECT, DEMO, DISC, or ICA. Student’s learning was subsequently assessed through six quizzes and four exams, which tested the constructs on three of Bloom’s cognitive levels: knowledge, comprehension, and application.

**Participants**

Participants, enrolled in the course during the spring semester at a Midwestern university agreed to participate in a study assessing various teaching strategies. The student body \((n = 51)\) composition consisted of 18 men and 33 women, with an average GPA of 3.31 \((SD = .66)\). However, two students were dropped from analyses due to incomplete records. The class mean age was 19.36, \(SD = .76\). The majority of students were Caucasian (58%), although other ethnicities were also represented: African-American (2%), Hispanic (8%), Asian (8%), Bi-Racial (4%) and other (4%). Additionally, almost half of students were psychology majors (46%), or double majoring in psychology (28%).

**Procedure**

Over the duration of the semester, constructs were taught by the instructor in ways that were complementary to the construct. In other words, if the instructor was unaware of a way to teach a construct through an in-class activity, it was not forced. For example, obedience was taught through a small demonstration, as opposed to the other methods, be-
cause it fit easily within the parameters of the classroom. A teaching assistant, who was blind to the hypotheses, was trained prior to the beginning of the semester to identify and code multiple teaching techniques. The constructs that were included in the analysis were based on the assistant’s notes. Then, two additional researchers, also blind to the hypotheses, created six quizzes, to be administered approximately every three weeks. Quizzes consisted of true/false, multiple choice, and short answer questions. For each construct on a quiz, three questions assessed learning: one question for each of the three levels of Bloom’s taxonomy (i.e. knowledge, comprehension, and application). Students were awarded one extra credit point for completing each quiz. In addition to the quizzes, four exams were created by the instructor and given to all of the students as part of their class grade. Finally, two additional researchers, who were also blind to the hypotheses, graded the quizzes. The exams were graded by the instructor.

**Measures**

For each quiz item, answers were either marked as completely correct or completely incorrect. Blank answers were graded as incorrect. No portions of credit were assigned. While the quizzes assessed learning on more constructs than just those posed in this current study, only quiz items pertaining to the current study were used in the analyses.

Exam grades were given as part of the normal class requirements. For multiple choice and fill-in-the-blank items, answers were graded as either completely correct or completely incorrect. For short answer and essay items, partial credit could be assigned. Blank answers were graded as incorrect.

**Results**

To examine the effectiveness of each teaching technique on each of the levels of assessment, four repeated measures ANOVAs were conducted examining differences between the three levels of Bloom’s taxonomy on items within the same teaching technique. A Bonferroni correction was applied to the probability at which the tests were accepted ($p < .017$). For each hypothesis, pairwise comparisons were analyzed for differences in the percentage of correct responses on test items.

The first hypothesis stated that LECT would be most effective for knowledge level assessments. A Mauchly’s test indicated that the assumption of sphericity had been violated ($\chi^2 = 9.66, p = .008$) therefore degrees of freedom was corrected using the Huynh-Feldt estimates of sphericity (epsilon = .871). The results indicate that there was a significant difference in the percentage of correct responses by assessment level ($F(1.74, 83.53) = 22.94, p = .000$). The percentage of correct scores on knowledge level assessments was significantly lower than both comprehension ($p = .000$) and application ($p = .000$). Thus, the first hypothesis was not supported. Refer to Table 1 (listed directly after the results of the fourth hypothesis) for a list of means and standard deviations for each technique on each level of Bloom’s taxonomy.
The second hypothesis stated that DEMO would be most effective for application level assessments. A Mauchly’s test indicated that the assumption of sphericity had been violated ($\chi^2 = 14.81, p = .001$) therefore degrees of freedom was corrected using the Huynh-Feldt estimates of sphericity (epsilon = .809). The results indicate that there was a significant difference in the percentage of correct responses by assessment level ($F(1.62, 77.70) = 4.64, p = .018$). Scores on application level assessments was significantly higher than knowledge ($p = .000$), but only marginally higher than comprehension ($p = .062$). Thus, the second hypothesis was only partially supported (see Table 1).

The third hypothesis stated that DISC would be most effective for comprehension level assessments. A Mauchly’s test indicated that the assumption of sphericity had been violated ($\chi^2 = 19.58, p = .000$) therefore degrees of freedom was corrected using the Greenhouse-Geisser estimates of sphericity (epsilon = .746). These results show that there was a significant difference in the percentage of correct responses by assessment level ($F(1.49, 71.61) = 28.60, p = .000$). Scores on comprehension level assessments were significantly lower than both knowledge ($p = .000$) and application ($p = .000$). Thus, the third hypothesis was not supported (see Table 1).

The fourth hypothesis stated that ICA would be most effective for comprehension and application level assessments. A Mauchly’s test indicated that the assumption of sphericity had not been violated ($\chi^2 = .56, p = .758$). Results indicate that there was a significant difference in the percentage of correct responses by assessment level ($F(2, 98) = 11.11, p = .000$). Scores on comprehension ($p = .000$) and application ($p = .007$) were both significantly higher than knowledge level scores. Thus, the fourth hypothesis was supported.

Table 1. Means and Standard Deviations of Scores for each Teaching Technique

<table>
<thead>
<tr>
<th></th>
<th>Knowledge</th>
<th>Comprehension</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>Lecture</td>
<td>.760**</td>
<td>.169</td>
<td>.878</td>
</tr>
<tr>
<td>Demonstrations</td>
<td>.678</td>
<td>.212</td>
<td>.698</td>
</tr>
<tr>
<td>Discussions</td>
<td>.820</td>
<td>.180</td>
<td>.621**</td>
</tr>
<tr>
<td>In-Class Activity</td>
<td>.789**</td>
<td>.163</td>
<td>.872</td>
</tr>
</tbody>
</table>

** = Was different from remaining Bloom’s levels ($p < .001$).

$^1$ = Was different from Knowledge ($p < .001$), and marginally significant from Comprehension ($p = .06$).

The fifth and final hypothesis stated that as the technique became increasingly active, so would the scores on test items. Thus, it was expected that students’ overall scores would be significantly higher for constructs taught using ICA than LECT methods. To analyze this, a repeated measure ANOVA was conducted using the teaching technique as the independent variable and the overall percentage of correct scores as the dependent variable.
Again, a Bonferroni correction was applied to the probability at which the post hoc comparisons were accepted ($p < .013$). A Mauchly’s test indicated that the assumption of sphericity had been violated ($\chi^2 = 33.28, p = .000$) therefore degrees of freedom was corrected using the Greenhouse-Geisser estimates of sphericity ($\varepsilon = .684$). These results indicate that there was a significant difference in the percentage of correct responses by assessment level ($F(1.96, 88.29) = 29.60, p = .000$). Scores for constructs taught using LECT ($M = .64, SD = .22$) was significantly lower than DEMO ($M = .79, SD = .14, p = .001$), DISC ($M = .82, SD = .11, p = .000$), and ICA ($M = .89, SD = .06, p = .000$). Moreover, scores on constructs taught using ICA were significantly higher than scores from LECT ($p = .000$), DEMO ($p = .000$), and DISC ($p = .005$). Thus, this hypothesis was supported. See Figure 1 for an illustration of scores for overall learning in each technique.

![Figure 1. Means and Standard Deviations for Overall Scores of each Technique](image)

**Discussion**

The current study sought to examine the effectiveness of four teaching techniques (i.e. lecture, demonstrations, discussions, and in-class activities) in the classroom. As each technique offers different benefits, the effectiveness of each technique was expected to vary by depth of learning on Bloom’s taxonomy (i.e. knowledge, comprehension, and application). The current findings indicate that each teaching technique has its own unique benefits and is effective for various types of learning.

Our first hypothesis, that lecture would be most effective on knowledge level questions was not supported. In fact, the lecture method was actually least effective as correct scores on knowledge level assessments were significantly lower than both comprehension and application. Perhaps this is because knowledge level assessments are often based on
rote memorization, such as knowing which definition describes a particular construct. Lecture, while it may contain explanations viable for comprehension level learning, and examples that are important for application skills, does not necessarily lend itself to increased memorization. In fact, some of the complaints about lecture are that students often seem bored, sleeping, or multi-tasking (Michel, et al., 2009; Van Eynde & Spencer, 1988). However, it should be noted that our analysis was specifically looking at differences between the three levels of Bloom’s taxonomy and did not compare lecturing to a control group, or not being introduced to the concept at all. The percentage of correct responses (76%) on knowledge level questions was still rather high. Thus, one could conclude that lecture is actually incredibly effective for all three levels of learning and perhaps these findings provides evidence contrary to the reputation that lecture has earned.

Our second hypothesis, that demonstrations in class would be most effective for comprehension, was also not supported. Instead, our findings indicate that demonstrations were no more effective for comprehension items than they were for knowledge level items. However, demonstrations were most effective for application level test items. This was somewhat surprising, as it was argued that while demonstrations increase attention and enjoyment, they only allow a few students from the classroom to actually manipulate and apply the information, leaving the rest of the students unengaged. However, it would appear that just watching others apply the information to a new situation is enough to learn the application oneself.

The findings from our third hypothesis, which stated that discussions would lead to higher answers on comprehension level questions, were possibly the most surprising. Our findings suggest that scores on comprehension were lower than both knowledge and application level items. While, Hadjioannou (2007) argued that students must first understand what another student has said in order to contribute to the discussion, this was not supported in our sample. In many cases, discussion could include thoughtful and thought provoking comments. However, they also include wrong thoughts, misleading information, and even mythology and urban legends. While the instructor takes this opportunity to correct and inform the students, perhaps the fact that the information has already been said is enough to ‘throw off’ some students, as it pertains to understanding. Recent research suggests that just seeing a wrong answer can interfere with one’s ability to learn the correct answer (Fazio, Agerwal, Marsh, & Roediger III, 2010; Roediger III and Marsh, 2005). Perhaps, in the case of comprehension, discussions including impromptu student explanations may actually be more hurtful than helpful. However, we did not examine this possibility. Future studies may want to empirically examine the effectiveness of varying types of discussions, specifically looking at examples when incorrect information is included in the discussion.

While, it was contrary to what was hypothesized, it should be noted that for the constructs taught using discussion, the percentage of correct responses on application and knowledge level items was above a more than satisfactory 80%. This lends evidence to the notion that allowing students to interact via discussions is an effective teaching technique. Perhaps as they repeatedly hear vocabulary words throughout the discussion, it
lends itself to increased memory. And, as students voice various personal stories involving the phenomena, others are able to learn how the phenomenon applies to multiple situations. But, once again, this was not empirically examined in the current study.

Our results did support our fourth hypothesis, which stated that in-class activities would increase scores on both comprehension and application level test items. However, it should be noted that while comprehension and application level test items were significantly higher than knowledge level items, the average score on knowledge level items was 79%. Arguably, this lends more evidence to the claims that in-class activities (ICA) are the most effective of all the techniques because they allow students to actually manipulate and practice applying the information for their selves (Forsyth, 2003; Whetten & Clark, 1996).

Finally, our findings supported the notion that active techniques do aid in increasing learning as in-class activities led to higher overall scores while lecture led to the lowest overall scores. However, this does not mean that one should blindly use active techniques in lieu of other methods. We often think of lecture and active teaching techniques as competing forces. This dichotomous thinking of good and bad techniques can be counterproductive. Even in the current study, no one method emerged as the ‘easy button’ of teaching or learning. Scores on quizzes and exams were fairly high, even when using lecture. It is unfortunate that lecture has earned such a bad reputation. While there may be some exceptions, the current research suggests that, in general, any technique that an instructor uses can be effective, if it is used competently, appropriately, and enthusiastically.

On the other hand, if your course is focused on decision making, rather than recalling facts, then active techniques probably should be a necessary component of your teaching repertoire (Serva & Fuller, 1998). As our findings suggest, active techniques affect learning on deeper levels. Additionally, active teaching can be an added bonus for teachers who are managing students with diverse learning styles. This is because instructors who vary their presentation methods create extra learning opportunities for students with different learning styles (Cook & Hazelwood, 2002). However, instructors should also realize that active teaching takes time away from full content coverage. Therefore, instructors should carefully evaluate whether using active techniques is worth sacrificing class time that could be used to cover other important information (Yoder & Hochevar, 2005).

Although, the current study is an important contribution to the literature in higher education there are limitations to the current study. One limitation to this study is that the students self-selected into the class. As this is an upper level psychology course, a large portion of the students (46%) were psychology majors, and were likely inherently interested in the material covered, and had received prior exposure to many of the topics in their introduction classes. Furthermore, the instructor tends to be liked among her students and many (42%) had taken a class with her previously. Perhaps many of the students were more alert, comfortable, and acclimated to the various teaching techniques being applied. Thus, the high scores could be explained through mere exposure or prior practice with both the phenomena and the format of the class. Since students self-selected
into the course, they could have chosen this instructor specifically because she uses many active techniques. If a student knows that they learn best when these techniques are used, and then subsequently chose to take a class from this instructor, the results could have been influenced in such a way that supports the hypotheses. However, this seems to be unlikely, considering that not all hypotheses were supported. Future studies should utilize these same methods with other instructors, as a means of replication.

Another limitation is the potential for experimenter bias. The data was collected in a classroom, and we attempted to control as many variables as possible without sacrificing the natural art of good teaching, but experimenter bias may have occurred as the experimenter was also the teacher. For example, constructs that lent themselves to being taught through in-class activities were used, and constructs for which the instructor had no knowledge of any interactive strategy were chosen for lecture. This could have led to more complicated or uninteresting constructs being taught through lecture, which could also have led to lower test grades. While efforts were made to minimize potential experimenter effects (e.g. coding constructs retroactively using TA’s notes, having multiple blind quiz makers and quiz graders), it is highly possible that some effects were left uncontrolled. Future studies should be aware of their occurrence and attempt to avoid error in the data.

Teaching is a complex endeavor. Combined factors, such as student motivation and the instructor’s rapport with the students, have the potential to influence how effective any technique is (Tomcho & Foels, 2008). Thus, any data taken from a classroom is inherently contaminated and may not provide a perfect picture of effectiveness. Regardless, based on the current study, active teaching techniques do enhance learning as quiz and exam scores were higher when students were allowed to interact with the material. While, results indicated that lecturing was the least effective technique, it should be noted that students still scored relatively high after lecturing alone, which indicates that learning was still occurring. Perhaps, most techniques are effective on some level, and the real decision should be on a construct by construct, and class session by class session basis. In the end, instructors must decide for themselves, and be confident in their decisions, regarding what techniques to use, what material to use it with, and how often to use them. That is probably the real underlying solution to the effectiveness of any teaching technique.

References


Evaluating and Improving the Assessment of Undergraduate Student Writing in a School of Business at a Large Regional University

David Glew, Tracy Meyer, Becky Sawyer, Pete Schuhmann, and Barry Wray

UNC Wilmington, Wilmington, North Carolina, 28403

Abstract

Business schools are often criticized for the inadequate writing skills of their graduates. Improving writing skills involves first understanding the current skill level of students. This research attempts to provide insights into the effectiveness of the current method of assessing writing skills in a school of business at a large regional university and to provide insights relative to how the assessment process can be improved. Three studies were conducted. Study 1 is a quantitative analysis of the full set of data based on the use of Criterion®, an online writing evaluation service developed by Educational Testing Service (ETS). Study 2 is a quantitative analysis of a coded subset of the data and Study 3 involves a qualitative data collection effort consisting of comments gathered from assessment proctors. While the overarching purpose of this research was to determine the effectiveness of the assessment tool utilized towards understanding the current status of students’ writing skills, the results also offer recommendations to improve the assessment process such that a clear picture of writing weaknesses can be more effectively addressed.

Keywords: Writing skills, assessment, teaching effectiveness.

Accreditation by the Association to Advance Collegiate Schools of Business (AACSB) is the highest certification a business school in the United States can attain. In order to achieve and maintain this accreditation a school must develop a systematic process to “develop, monitor, evaluate, and revise the substance and delivery of the curricula of degree programs and to assess the impact of the curricula on learning.” A very important component of this curriculum is student communication abilities. Therefore business schools wishing to be accredited by the AACSB must ensure all students acquire the necessary skills to be effective oral and written communicators. In order to accomplish the goal of ensuring effective writing skills, one business school at a large regional university uses an automated online evaluation system that requires students to write a timed essay and then provides a holistic score as well as sub-scores for grammar, usage, mechanics and style for each student. This research examines the effectiveness of this online student writing skills assessment process and provides recommendations for improving the process.

The business school has stated as one of its undergraduate learning goals:

1 Corresponding author's email: wrayb@uncw.edu

©2011 All rights reserved
“Our students will be able to conceptualize a complex issue into a coherent written statement and oral presentation, demonstrated with the effective use of technology.”

To measure the extent to which students are able to conceptualize a complex issue into a coherent written statement, the following process was implemented:

“To effectively measure written communication, the business school has adopted an ETS instrument developed through AACSB. This instrument is administered in the capstone classes of each major every semester. Faculty are trained in administration of the writing assessment.”

This learning goal and corresponding assessment process are part of the broader business school Assurance of Learning Plan. Criterion®, an online writing evaluation service developed by Educational Testing Service (ETS), was selected to measure students’ writing proficiency. Criterion® requires students to write a timed essay on a given subject which is then evaluated by the ETS computerized e-rater® scoring engine and Critique™ writing assessment tools. Business school writing assessments began in spring 2008 and have been conducted in each department’s capstone courses. As of the start of fall 2009, approximately 800 students had completed the writing assessment.

A research team composed of one member from each department of the business school was formed to analyze the results to date from the Criterion® assessments. Specifically, the research team was charged with “making the ‘first cut’ at interpreting the data, focusing on the areas that should be improved and identifying ways they can be improved.”

**Research Question**

The question for the research reported in this article is “how effective is the current method of assessing writing skills in the business school?” Three studies were conducted to analyze the Criterion® assessment data: 1) a quantitative analysis of the full set of Criterion® data, 2) a quantitative analysis of a coded subset of data, and 3) a qualitative analysis of comments gathered from assessment proctors.

**Literature Review**

The current research on assessment of writing skills in business schools is limited. Warnock (2009) described a pilot effort for an accreditation-driven writing assessment in a business school and detailed the pilot’s logistics and methods. Bacon, Paul, Johnson and Conley (2008) and Bacon and Anderson (2004) researched assessing and improving writing skills of marketing students by using an internally developed assessment tool. They developed a computer-scored measure of students’ ability to edit for writing mechanics. The authors tested the method with a pretest and posttest design and found that when stu-

---

2 Criterion® and e-rater® are registered trademarks of Educational Testing Service (ETS). Critique™ is a trademark of ETS.

*The Journal of Effective Teaching, Vol. 11, No. 2, 2011, 55-74* 
©2011 All rights reserved
students were held accountable for improvement (with 5% of the course grade dependent on improvement) there was a significant increase in students’ ability to edit. Without that accountability, improvement did not occur. Wolf, Connelly, and Komare (2008) and Wolf and Stevens (2007) examined the role of rubrics in assessing teaching and student learning. They found different assessment rubrics can be used to improve both courses and programs in different fields and disciplines. In addition, Quible, Zane, and Griffin (2007) examined possible reasons for writing deficiencies for business writers. They offer evidence that a modified context-based approach, the glossing approach, and consistent error marking can reduce the number of sentence-level errors students make. Fraser et al. (2005) developed a multilevel assessment of business writing using both an online objective test and in class writing exercise. Zhao and Alexander (2004) identified short- and long-term impacts of business communication education on students’ writing skill developments and performance outcomes. Earlier research in assessing business writing was done by Varner, Iris, and Pomereneke (1998). They determined instructors need to be aware of issues such as developing clear goals from the beginning to achieve effective reliability and validity when conducting a writing assessment.

Methodology

Two studies were conducted to analyze the Criterion assessment data:

- Study 1: A quantitative analysis of the full set of Criterion data
- Study 2: A quantitative analysis of a coded subset of data

In addition, a third study was conducted to more closely examine the assessment process:

- Study 3: A qualitative analysis of comments gathered from assessment proctors

Study 1

Sample. Over the period of spring 2008 to summer 2009, 860 business school seniors took the Criterion writing assessment in their capstone course. Due to incomplete data, 37 observations were dropped leaving a sample size of 823.

Assessment. Upon logging in to the Criterion online assessment system, students were given these instructions: “Present your perspective on [the following issue], using relevant reasons and/or examples to support your views. ‘It is important for higher education to challenge established traditions and values’” (italics added). Students were required to submit their essay within 45 minutes.

Variables. The Criterion assessment data includes the following variables (a detailed explanation of these variables is provided in Appendices A and B):

1. Time to complete the essay
2. Holistic score (ordinal scale from low of 1 to high of 6)
3. Number of grammar errors
4. Number of usage errors
5. Number of mechanics errors
6. Number of style comments

Analyses. Descriptive statistics and inter-item correlations are provided in Tables 1 and 2. For ease of interpretation, the distribution of holistic scores is depicted in Figure 1. The mean and standard deviation of the number of errors or comments for grammar, usage, and mechanics is listed in Table 1. To further illustrate the prevalence or absence of errors or comments, the number of essays with no errors or comments in each category is shown in Figure 2.

Table 1. Study 1 Descriptive Statistics

<table>
<thead>
<tr>
<th></th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holistic Score</td>
<td>1</td>
<td>6</td>
<td>3.77</td>
<td>1.10</td>
</tr>
<tr>
<td>Time (minutes)</td>
<td>.03</td>
<td>45.00</td>
<td>29.49</td>
<td>11.26</td>
</tr>
<tr>
<td>Grammar Errors</td>
<td>0</td>
<td>19</td>
<td>2.50</td>
<td>2.31</td>
</tr>
<tr>
<td>Usage Errors</td>
<td>0</td>
<td>12</td>
<td>2.08</td>
<td>1.88</td>
</tr>
<tr>
<td>Mechanics Errors</td>
<td>0</td>
<td>57</td>
<td>1.57</td>
<td>3.53</td>
</tr>
<tr>
<td>Style Comments</td>
<td>0</td>
<td>89</td>
<td>13.36</td>
<td>14.50</td>
</tr>
</tbody>
</table>

Figure 2. Study 1 Student Essays with No Errors or Style Comments

A more in-depth investigation of these data was conducted by performing an ordered logit multiple regression analysis of holistic score on other variables (see McKelvey & Zavoina, 1975 for details). Since essays would likely contain more or fewer errors simply because of their length (e.g., longer essays would be expected to have more errors) variables 3 – 6 above were standardized by dividing the number of errors or comments in
each category by the time taken to complete the essay. Hence three “errors per minute” and one “comments per minute” variable were created.

The ordered logit multiple regression analysis produced the following results:

1. While controlling for major (using 7 indicator variables with POM as the baseline), the number of grammar errors per minute increases the holistic score. This result is highly statistically significant ($p$-value < 0.001).

2. While controlling for major (using 7 indicator variables with POM as the baseline), the number of usage and mechanics errors and style comments per minute decreases the holistic score. These results are highly statistically significant ($p$-values < 0.005).

These results must be viewed with caution however due to the existence of multicollinearity in the independent variables (Table 2).

**Table 2. Study 1 Inter-item Correlations**

<table>
<thead>
<tr>
<th></th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Holistic Score</td>
<td>.45**</td>
<td>.31**</td>
<td>.27**</td>
<td>-.05</td>
<td>-.13**</td>
</tr>
<tr>
<td>2. Time</td>
<td>–</td>
<td>.20**</td>
<td>.29**</td>
<td>.14**</td>
<td>-.12**</td>
</tr>
<tr>
<td>3. Grammar Errors</td>
<td>–</td>
<td>.20**</td>
<td>.33**</td>
<td>.37**</td>
<td>.09*</td>
</tr>
<tr>
<td>4. Usage Errors</td>
<td>–</td>
<td>–</td>
<td>.19**</td>
<td>-.03</td>
<td></td>
</tr>
<tr>
<td>5. Mechanics Errors</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td></td>
<td>.06</td>
</tr>
<tr>
<td>6. Style Comments</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>–</td>
</tr>
</tbody>
</table>

*Correlation is significant at the 0.05 level (2-tailed).
**Correlation is significant at the 0.01 level (2-tailed).

**Study 2**

**Sample.** A stratified random sample of 119 essays was drawn from the larger data set of 823 essays used in study 1. Twenty essays for each holistic score were included (except for holistic score 6 for which only 19 essays were available).

**Coding and Variables.** The same error and comment variables described in study 1 were available for the 119 essays. In addition, variables related to the length and content of the essays were calculated (e.g., key words from the essay topic "It is important for higher education to challenge established traditions and values" were counted). The variables that were calculated were:

1. Number of words
2. Average word length
3. Number of sentences
4. Average sentence length
5. Instances of the phrase “higher education”
6. Instances of the word “challenge”
7. Instances of the word “established”
8. Instances of the word “values”
9. Instances of the word “tradition”
10. Instances of the word “importance”

**Analyses.** Descriptive statistics and inter-item correlations are provided in Tables 3 and 4.

**Table 3. Study 2 Descriptive Statistics**

<table>
<thead>
<tr>
<th></th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holistic Score</td>
<td>1</td>
<td>6</td>
<td>3.48</td>
<td>1.71</td>
</tr>
<tr>
<td>Grammar Errors</td>
<td>0</td>
<td>14</td>
<td>2.69</td>
<td>2.87</td>
</tr>
<tr>
<td>Usage Errors</td>
<td>0</td>
<td>9</td>
<td>1.87</td>
<td>1.85</td>
</tr>
<tr>
<td>Mechanics Errors</td>
<td>0</td>
<td>43</td>
<td>1.85</td>
<td>4.62</td>
</tr>
<tr>
<td>Style Comments</td>
<td>0</td>
<td>89</td>
<td>16.64</td>
<td>15.13</td>
</tr>
<tr>
<td>Total Words</td>
<td>53</td>
<td>1057</td>
<td>364.96</td>
<td>243.94</td>
</tr>
<tr>
<td>Avg. Word Length</td>
<td>4.11</td>
<td>5.94</td>
<td>4.99</td>
<td>.34</td>
</tr>
<tr>
<td>Total Sentences</td>
<td>3</td>
<td>53</td>
<td>17.60</td>
<td>10.96</td>
</tr>
<tr>
<td>Avg. Sentence Length</td>
<td>10.33</td>
<td>39.15</td>
<td>20.79</td>
<td>4.76</td>
</tr>
<tr>
<td>“Higher Education”</td>
<td>0</td>
<td>14</td>
<td>4.19</td>
<td>3.14</td>
</tr>
<tr>
<td>“Challenge”</td>
<td>0</td>
<td>20</td>
<td>4.09</td>
<td>3.12</td>
</tr>
<tr>
<td>“Established”</td>
<td>0</td>
<td>10</td>
<td>2.60</td>
<td>2.45</td>
</tr>
<tr>
<td>“Values”</td>
<td>0</td>
<td>17</td>
<td>5.73</td>
<td>3.83</td>
</tr>
<tr>
<td>“Traditions”</td>
<td>0</td>
<td>19</td>
<td>4.96</td>
<td>3.35</td>
</tr>
<tr>
<td>“Important”</td>
<td>0</td>
<td>9</td>
<td>1.88</td>
<td>1.55</td>
</tr>
</tbody>
</table>

An ordered logit multiple regression of holistic score on grammar, usage, mechanics and style errors per word was conducted. In study 1, grammar, usage, mechanics and style variables were standardized to account for different amounts of time spent on the essays. In study 2, a similar standardization was performed to convert these values into three “errors per word” and one “comments per word” variables.

The ordered logit multiple regression analysis in study 2 produced the following results:

1. The coefficient on grammar errors per word was *positive* and highly statistically significant (i.e., higher holistic scores were associated with more grammar errors).
2. The coefficients on mechanics and style comments per word were *negative* and statistically significant (i.e., higher holistic scores were associated with fewer mechanics errors and fewer style comments).
3. The coefficient on usage errors per word was not statistically significant.

These results must be viewed with caution however due to the existence of multicollinearity in the independent variables (Table 4).

**Study 3**

*Sample.* The sixteen business school faculty members who proctored the *Criterion* assessment in their courses were asked to briefly respond the questions shown below. Eleven (69%) faculty members responded.

*Qualitative Content Analysis.* The eleven responses to each question were examined for similarities and differences. The summaries of these responses are based on the subjective judgment of the research team.

**Q1. How do you fit the assessment into your course (e.g., set aside one day of the semester, ask students to take the assessment outside of class)?**

Four (36%) respondents used class time to give the assessment. Seven (64%) asked students to take the assessment outside of class.

**Q2. Do you make the assessment a course requirement, give extra credit, or in any other way link the assessment to students’ course grades? Do you use any other incentives?**

Five (45%) respondents made the assessment a course requirement, but include no other incentives. Three (27%) respondents linked taking the assessment to the students’ course grades, but did so variously by offering a nominal amount of extra credit or telling students they would lose points if they did not complete it. Two respondents (18%) reported giving extra credit.

Overall, responses to this issue varied from providing incentives to separating the assessment as much as possible from the course. In addition, some respondents approached the assessment differently at the time they answered the questionnaire than they had in the past. At least one respondent mentioned that the writing assessment was unpopular with students and had negatively affected course evaluations.

**Q3. Besides the technical information of how to login to the ETS *Criterion* website, what other information do you give the students? For example, do you explain the purpose of the assessment? Do you play up or downplay the importance of the assessment? Do you encourage a high level of effort or do you try to remain neutral and let students determine their own level of effort?**

Most respondents encouraged students to do their best. They explained the importance of the assessment for business school accreditation purposes, and highlighted the potentially
Table 4. Study 2 Inter-item Correlations

<table>
<thead>
<tr>
<th></th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Holistic Score</td>
<td>.46 **</td>
<td>.41 **</td>
<td>-</td>
<td>.14</td>
<td>.91 **</td>
<td>-</td>
<td>.04</td>
<td>.85 **</td>
<td>.27 **</td>
<td>.51 **</td>
<td>.52 **</td>
<td>.48 **</td>
<td>.58 **</td>
<td>.48 **</td>
</tr>
<tr>
<td>2. Grammar Errors</td>
<td>–</td>
<td>.47 **</td>
<td>.38 **</td>
<td>.36 **</td>
<td>.60 **</td>
<td>-</td>
<td>.39 **</td>
<td>.48 **</td>
<td>.35 **</td>
<td>.41 **</td>
<td>.51 **</td>
<td>.47 **</td>
<td>.48 **</td>
<td>.44 **</td>
</tr>
<tr>
<td>3. Usage Errors</td>
<td>–</td>
<td>.34 **</td>
<td>-</td>
<td>.03</td>
<td>.46 **</td>
<td>-</td>
<td>.01</td>
<td>.40 **</td>
<td>.20 **</td>
<td>.32 **</td>
<td>.39 **</td>
<td>.22 **</td>
<td>.33 **</td>
<td>.30 **</td>
</tr>
<tr>
<td>4. Mechanics Errors</td>
<td>–</td>
<td>.27 **</td>
<td>-</td>
<td>.02</td>
<td>-</td>
<td>.25 **</td>
<td>-</td>
<td>.14</td>
<td>-</td>
<td>.02</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5. Style Comments</td>
<td>–</td>
<td>-</td>
<td>-</td>
<td>.04</td>
<td>-</td>
<td>.34 **</td>
<td>-</td>
<td>.01</td>
<td>-</td>
<td>.08</td>
<td>.00</td>
<td>-</td>
<td>.10</td>
<td>-</td>
</tr>
<tr>
<td>6. Total Words</td>
<td>–</td>
<td>-</td>
<td>-</td>
<td>.91 **</td>
<td>.33 **</td>
<td>.51 **</td>
<td>.49 **</td>
<td>.44 **</td>
<td>.56 **</td>
<td>.45 **</td>
<td>.41 **</td>
<td>.10</td>
<td>-</td>
<td>1.0</td>
</tr>
<tr>
<td>7. Avg. Word Length</td>
<td>–</td>
<td>-</td>
<td>-</td>
<td>.27 **</td>
<td>.01</td>
<td>.00</td>
<td>-</td>
<td>.25 **</td>
<td>.03</td>
<td>.10</td>
<td>-</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>8. Total Sentences</td>
<td>–</td>
<td>-</td>
<td>-</td>
<td>.02</td>
<td>.44 **</td>
<td>.51 **</td>
<td>.41 **</td>
<td>.53 **</td>
<td>.45 **</td>
<td>.38 **</td>
<td>-</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>9. Avg. Sentence Length</td>
<td>–</td>
<td>.21</td>
<td>.08</td>
<td>.11</td>
<td>.13</td>
<td>.06</td>
<td>.14</td>
<td>-</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>10. Higher Education</td>
<td>–</td>
<td>.39 **</td>
<td>.39 **</td>
<td>.40 **</td>
<td>.29 **</td>
<td>.41 **</td>
<td>-</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>11. Challenge</td>
<td>–</td>
<td>.45</td>
<td>.54 **</td>
<td>.60 **</td>
<td>.43</td>
<td>-</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>12. Established</td>
<td>–</td>
<td>.62 **</td>
<td>.59 **</td>
<td>.39 **</td>
<td>-</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>13. Values</td>
<td>–</td>
<td>.83 **</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>14. Traditions</td>
<td>–</td>
<td>.42 **</td>
<td>-</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>15. Important</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>

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

valuable diagnostic feedback students would receive. Other respondents remained neutral or informed students that they would review the time students spent writing their essay and base course credit on their apparent effort.

Q4. What is your impression of how seriously the students take the assessment?

Responses to this question varied widely. Some felt most students took the assessment seriously because they were interested, wanted to get a good score, or wanted the busi-
ness school to look good. Others stated that students did not take the assessment seriously because of a lack of consequences or because the essay question was too abstract. Some respondents believed their efforts at encouraging students, including changing the way they included the assessment in their courses, had increased students’ motivation.

Q5. In your opinion, is the ETS Criterion online writing assessment a good way to measure writing ability? Is there a different approach you would suggest we take?

Faculty members were divided on whether Criterion is a good way to measure writing ability. Some pointed to efficiency and consistency as the primary advantages of an automated scoring system. Other respondents noted that the system could not evaluate content, so students could write their essay on an unrelated topic and still receive good scores. This was especially true given the topic students were asked to address. The lack of comparison data from a national sample was also lamented. A third group of respondents were simply unsure whether Criterion was a good way to assess writing ability.

Recommendations for alternative approaches included offering a business communications course, hiring a trained/professional writer to evaluate students’ work, and assessing writing within a course based on existing or additional written assignments.

Respondents were invited to make other comments

Additional comments included:

- Recommending and offering support for a business communications class
- Recommending a stronger focus on writing in the freshman English course
- Reporting an increased focus on writing within the respondents’ own courses
- Requiring students to re-write poor quality work and to visit the writing center
- Asking where business school students actually learn to write
- Noting that writing ability should have been addressed in admissions testing and in English courses

Interpretation of Results

Quantitative Data Analyses

The business school has adopted the following self-defined labels for three categories of holistic scores:

<table>
<thead>
<tr>
<th>Holistic Score</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-2</td>
<td>Needs a lot of work</td>
</tr>
<tr>
<td>3-4</td>
<td>Needs some work</td>
</tr>
<tr>
<td>5-6</td>
<td>Doing ok</td>
</tr>
</tbody>
</table>
At first glance, the quantitative data analyses suggest business school students, with an average holistic score of 3.77, need some improvement in their writing, but not much. Average error rates for grammar (2.50 errors per essay), usage (2.08) and mechanics (1.57) do not appear to indicate a crisis-level situation. Moreover, a substantial proportion of students made no errors in grammar (16.65% made no grammar errors), usage (21.26%), and mechanics (43.86%).

A more in-depth analysis of the data reveals several unexpected results. Holistic score had a strong significant positive correlation with grammar and usage errors. In fact, grammar errors were the best predictors of holistic scores in the ordered logit regression analyses. In other words, the essays receiving the highest holistic scores were the ones with the most grammar errors. However, there was multicollinearity in the logistic regression model as a result of strong correlations between the independent variables (tables 2 and 4). The existence of multicollinearity inflates the variances of the parameter estimates. The inflated variances may result in lack of statistical significance of some individual independent variables while the overall model was strongly significant. Multicollinearity may also result in wrong signs and magnitudes of regression coefficient estimates and, consequently, incorrect conclusions about relationships between independent and dependent variables.

These unexpected results may also be explained in part by examining the relationship between holistic scores and two measures of essay length: time spent on the essay and number of words. Of these two measures, number of words is more accurate because some students wrote their essays outside of Criterion, and then cut and pasted the text into the online system (thus the recorded time spent by some students of mere seconds). Holistic score and time spent on the essay are correlated at $r = .45$. Holistic score and number of words are correlated at $r = .91$. This means the length of the essay alone explains approximately 83% of the variance in holistic scores.

Taken together, these relationships indicate that students who wrote longer essays made more grammar and usage errors, but also received higher holistic scores. One tentative conclusion that could be drawn from these results is that the Criterion evaluation system appears to give more weight to the length of an essay than to the appropriate use of grammar.

Style comments had a significant but negative association with holistic score. Criterion does not assign style “errors”, but rather counts items such as the repetition of certain words and sentences beginning with coordinating conjunctions. Each instance of these items creates a style comment (e.g., if a student used the word “we” fifteen times, the essay would receive fifteen style comments). Style comments appear more subjective than the grammar, usage and mechanics errors identified by Criterion, and do not necessarily indicate a problem with the student’s writing. For instance, the essay that received the highest number of style comments – 89 – also received the highest holistic score possible – 6. Due to these factors, style comments may offer little benefit in assessing student writing.
Mechanics errors also had a significant negative association with holistic score. However, nearly one-half of all students’ essays (43.86%) had no mechanics errors, so less improvement appears needed in this area than in the other error categories.

Holistic score was significantly correlated with the incidence of all six key words taken from the essay prompt: “higher education,” “challenge,” “established,” “tradition,” “values,” and “important.” The average correlation between holistic score and these words was $r = .49$.

**Qualitative Data Analysis**

A fundamental tenet of scholarly research is that data are no better than the procedure used to gather them. Comments by the faculty members who proctored the *Criterion* assessment reveal substantial differences in the way the assessment was administered. Across the various capstone courses, students received inconsistent instructions regarding the importance of the test and inconsistent incentives for successful completion. This is the case across proctors in a given semester as well as for some proctors across time.

Such inconsistency in the administration of the assessment creates a situation where otherwise strong writers could produce low-scoring essays. This induced variability in the variable we are attempting to explain (holistic score) makes any analysis of cause-and-effect suspect and likely unreliable.

Moreover, because the *Criterion* system grades essays using strictly mechanical means (word counts, error counts, etc) it is also highly likely that otherwise nonsensical essays could produce high holistic scores. This possibility was confirmed by a quick review of some of the actual essays in the dataset. They frequently drifted off topic and sometimes were completely unrelated to the essay prompt (e.g., one was written on Kobe Bryant and the LA Lakers; another begins “I really do not want to do this. This is probably the dumbest thing I have ever heard of.”; a third ends with “I love you.”).

ETS acknowledges these sorts of weakness in *Criterion*:

> **Can students trick the *Criterion* service?**  
> Yes. Since the e-rater engine cannot really understand English, it will be fooled by an illogical, but well-written, argument.”

> **Will the *Criterion* service catch cheating or plagiarism?**  
> No. The *Criterion* service simply evaluates the essay.”

In short, interpreting the results of the quantitative analyses described is severely limited by the variability in how the writing assessment has been administered thus far in the business school, and the inability of *Criterion* to accurately recognize the content of the essays. At best, these results should be interpreted with a great deal of caution.
Recommendations

Two general sources of information about business school student writing exist: 1) faculty anecdotes and 2) the Criterion assessment data addressed in this report. Faculty anecdotes bring to mind egregious instances of poor student writing, but are not a systematic evaluation of student skills. The Criterion assessment system comes closer to a systematic evaluation, but the varied procedures by which it has been administered severely compromises the reliability of its results.

These two sources of information about student writing, however, raise at least six issues (listed in Table 5). Each issue (column A), in turn, is related to a potential problem (column B), which then requires one or more solutions (column C). With the exception of inadequate assessment (which has been addressed to some extent in this report) the problems are described as “potential” because they have yet to be systematically documented in this sample of students. However, anecdotal evidence suggests they do exist.

The issues, problems and categories of solutions listed in Table 5 are not mutually exclusive, but they are arranged in the sequence by which the research team recommends they be addressed. For example, the problems of inadequate assessment and apathy must be addressed before it will be possible to accurately determine whether low skill level is truly a problem.

Table 5. Issues, Problems and Solutions

<table>
<thead>
<tr>
<th>A. Writing Issues</th>
<th>B. Potential Problems</th>
<th>C. Categories of Solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Skill assessment</td>
<td>Inadequate assessment</td>
<td>Improve assessment</td>
</tr>
<tr>
<td>2. Motivation</td>
<td>Apathy</td>
<td>Increase motivation</td>
</tr>
<tr>
<td>3. Skill level</td>
<td>Low skill level</td>
<td>Develop skills</td>
</tr>
<tr>
<td>4. Recency of skill acquisition</td>
<td>Old or stale skills</td>
<td>Refresh skills</td>
</tr>
<tr>
<td>5. Feedback frequency</td>
<td>Infrequent feedback</td>
<td>Provide frequent feedback</td>
</tr>
<tr>
<td>6. Practice frequency</td>
<td>Few opportunities for practice</td>
<td>Increase opportunities for practice</td>
</tr>
</tbody>
</table>

The research team’s strongest recommendations at this point focus on improving the assessment of student writing. However, recommendations related to the other issues have also been made.

**Improving Assessment**

**Recommendation 1: Standardize the assessment procedure.**

The primary weakness in the current assessment procedure is that individual faculty members take different approaches to administering Criterion. To improve reliability, the procedure should be identical for all students. This means all students who complete the
Criterion assessment should receive the same instructions, take the assessment under the same physical conditions, use the same word processing software, have the same requirement to take the assessment, and have the same incentive to do so.

The research team recommends that an assessment team (possibly comprised of research team members), rather than course instructors, be in charge of administering the assessment. This approach will standardize the instructions students receive. One class session during the semester should be designated for the assessment, and students should meet in the computer lab on that day. Students should write their essays entirely in Criterion rather than cutting and pasting them from other programs. Writing the essay entirely in Criterion has at least three advantages: it provides a better assessment of whether students have mastered writing skills (vs. rely on technology to catch errors); it reduces the chance that students will paste in an essay they did not write; and it reflects the actual time students spend writing the essay – a variable that may prove valuable in future analyses. Completing the assessment should be a requirement (for the course or for graduation) for all students. Any additional incentives (e.g., any impact on a course grade) should be the same for all students.

Some instructors’ are legitimately concerned that the assessment may have a negative impact on students’ perceptions of their teaching, so consideration should be given to ways the assessment might be disconnected from the capstone courses. As one possibility, the capstone courses could simply serve as a tracking mechanism to verify that students have completed the assessment. An “incomplete” grade might be assigned to students who do not complete the assessment, but there would be no other impact on the course grade.

**Recommendation 2: Add a human evaluator.**

The primary limitation of Criterion is that it cannot distinguish whether the content of an essay is related to the designated essay prompt. Adding a human evaluator who, at a minimum, simply determines whether the essays are on topic will substantially enhance the validity of the assessment. This arrangement is noted by ETS:

“How do institutions use the Criterion service for assessment purposes? Some institutions use the Criterion scores for exit testing — combining a Criterion score with the score from a faculty reader....”

“Educators can stop students from deliberately trying to fool the Criterion service by announcing that a random sample of essays will be read by independent readers.”

A human evaluator could also overcome other limitations in Criterion. For instance, twenty students have taken the assessment but have not received scores because Criterion was unable to identify certain organizational elements, such as thesis statements and conclusions. Upon closer review, these elements were “missing” because the students had
failed to double-space between paragraphs. A more valuable contribution of a human evaluator would be to provide insights related to specific writing skills that appear weak.

Using a human evaluator in addition to Criterion would raise the cost of the assessment, and would provide at least some redundant information. Consideration should be given, however, to the possibility of using a human evaluator instead of Criterion. A qualified graduate student, faculty member, or writing professional could provide much richer information regarding students’ writing skills, at a potentially lower cost. Developing uniform scoring guidelines would eliminate some, but not all, of the variability across human evaluators.

**Increasing Motivation**

**Recommendation 3: Create incentives for high performance.**

Currently, students have little, if any, external incentive to do well on the assessment. Adding a positive incentive to demonstrate a high level of writing ability (e.g., a holistic score of 5 or better) will result in a more accurate assessment of students’ true ability levels. Possible incentives include extra credit, a certificate and corresponding resume line item, a “Dean’s List”, a scholarship for the best essay, entry into a drawing for a prize such as an iPhone, and countless others. Possible negative incentives include an incomplete grade in the course, delayed graduation, required remedial instruction, or a requirement to re-take the assessment until a higher score is achieved.

**Recommendation 4: Develop an intrinsically interesting essay prompt.**

Students and faculty have commented that the current essay prompt of the role of higher education in modern society is too abstract and reduces students’ motivation. A more intrinsically motivating prompt, such as “How will graduating with a business school degree affect your future career?” will create greater focus on the question and ownership of the answer. The possibility of developing a customized essay prompt is outlined on the ETS website:

“...when educators want students to write on a topic not available in the Criterion library, they can create and assign their own prompt for a student assignment. Although essays written on educator-created topics do not receive the holistic score, all of the features of diagnostic feedback will be reported when the essay is submitted. Colleges and universities can also work with ETS to create new topics tailored to their needs.”

**Developing and Refreshing Skills**

**Recommendation 5: Include systematic writing instruction as part of the business school curriculum.**
One of the most commonly suggested ways to improve student writing is to add a business communications course to the business school core. The research team recognizes that developing and offering such a course would require substantial resources – ones that appear unavailable, at least in the foreseeable future. However, teaching writing skills is not an all-or-nothing proposition, and there is no reason to wait until a business communications course is a viable option to begin.

The research team does not make a formal recommendation here as to how such instruction should be provided. The appropriate level of instruction depends on a more accurate assessment of students’ existing skills than is currently available. In the meantime, the following ideas are offered to promote future discussions on how writing instruction might be implemented in the business school:

- Add a one-credit writing skills lab to the existing business school writing intensive courses
- Teach core writing skills modules in one or more sessions of the business school writing intensive courses, by the instructor or by a guest
- Develop a required online writing skills course, or adopt an existing course
- Promote self-study by requiring a minimum score on a standardized writing exam (e.g., a test similar to the GMAT verbal section) prior to admission to the business school
- Encourage passive learning by posting grammar questions/answers and “tips of the day” on the hallway monitors

**Providing Frequent Feedback and Increasing Opportunities for Practice**

The number of opportunities students have for deliberate, informed writing practice reflects the value of effective writing in the business school culture. If at the core of this culture the assumption is “writing is their problem” (“they” being the students, the English Department, the secondary education system, or anything else) then the level of writing feedback and the number of opportunities for practice are unlikely to change. If, in contrast, the core belief is “it is our responsibility to produce effective writers,” (“we” being the business school faculty) then steps will be taken to provide students with meaningful feedback and practice. No specific solution is recommended in this category; however, all organizations do well to understand their own culture.

Here are a few ways to increase feedback and opportunities for practice:

- Assign more short writing assignments rather than one long assignment
- Require first drafts and subsequent revisions
- Develop a writing scoring rubric to simplify grading. Potentially, a business school-wide scoring rubric could be developed that includes the most common errors in student writing. A companion document could be distributed to students.
- Utilize qualified business school graduate students (or graduate students outside business school) to assist with scoring grammar, usage and mechanics in written assignments. The instructor could still grade for content.
• Establish a business school writing lab

Conclusions

There is no doubt effective writing skills are among the most valuable tools business school students can take with them when they graduate. Assessment is a key to determining the current level of students’ writing skills. However, assessment alone leaves several key questions unanswered.

The question of “Why?”
Assessment alone does little to improve students’ writing skills. The Criterion system provides students with a limited amount of feedback on a short essay (an average length of eighteen sentences), but no chance to make revisions, and in practical terms, no incentive to even review the feedback. Improving the assessment of student writing (the principal recommendations of this report), although important, will do little to directly improve students’ skills.

The questions of “Where?” and “When?”
Measuring whether students are able to “conceptualize a complex issue into a coherent written statement” is necessary to assess where the business school stands relative to its formal learning goal. Although an attempt is being made to assess writing skills, these skills are not formally taught in business school courses. Where are they taught? One common assertion is that students should have these skills before they enter the business school. Anecdotal evidence suggests at least some students arrive at the business school without these skills.

Moving the assessment to the beginning of students’ coursework, instead of, or in addition to, performing an assessment at the end of their program would allow time for skill deficiencies to be corrected. Some of the recommendations given in this report suggest deficiencies can be corrected in the business school. Alternatively, a minimum level of writing proficiency could be added as a business school admission requirement (a practice not uncommon at other universities). In that case, skill deficiencies could be addressed outside the business school.

The question of “Who?”
Does measuring progress towards the business school learning goal require an assessment of all students? If no individual intervention is planned for low performers, assessing the full population of business school students is an expensive endeavor. The average level of student writing skill can be measured with a much smaller sample. Likewise, if not every student needs remedial writing instruction, requiring all students to participate in a skill development program would be an inefficient use of their time.

This article highlights some of the challenges one business school encountered while addressing the matter of student writing. Several recommendations are offered, but a variety of questions have yet to be answered. To be sure, others across universities face similar
issues. We look forward to further investigation and insight on this important and complex topic.

References


Appendix A: Criterion Diagnostic Feedback Categories

<table>
<thead>
<tr>
<th>Grammar Errors</th>
<th>Style Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>• fragment or missing comma</td>
<td>• repetition of words</td>
</tr>
<tr>
<td>• run-on sentences</td>
<td>• inappropriate words or phrases</td>
</tr>
<tr>
<td>• garbled sentences</td>
<td>• sentences beginning with coordinating</td>
</tr>
<tr>
<td>• subject-verb agreement</td>
<td>conjunctions</td>
</tr>
<tr>
<td>• ill-formed verbs</td>
<td>• too many short sentences</td>
</tr>
<tr>
<td>• pronoun errors</td>
<td>• too many long sentences</td>
</tr>
<tr>
<td>• possessive errors</td>
<td>• passive voice</td>
</tr>
<tr>
<td>• wrong or missing word</td>
<td></td>
</tr>
<tr>
<td>• proofreading errors</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Usage Errors</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>• wrong article</td>
<td>• Thesis statement</td>
</tr>
<tr>
<td>• missing or extra article</td>
<td>• Main ideas</td>
</tr>
<tr>
<td>• confused words</td>
<td>• Supporting ideas</td>
</tr>
<tr>
<td>• wrong form of word</td>
<td>• Conclusion</td>
</tr>
<tr>
<td>• faulty comparisons</td>
<td>• Introductory material</td>
</tr>
<tr>
<td>• preposition error</td>
<td>• Other</td>
</tr>
<tr>
<td>• nonstandard word form</td>
<td>• Transitional words and phrases</td>
</tr>
<tr>
<td>• negation error</td>
<td>• Repetition of ideas</td>
</tr>
<tr>
<td></td>
<td>• Topic relationships and technical</td>
</tr>
<tr>
<td></td>
<td>quality</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mechanics Errors</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>• spelling</td>
<td>• Thesis statement</td>
</tr>
<tr>
<td>• capitalize proper nouns</td>
<td>• Main ideas</td>
</tr>
<tr>
<td>• missing initial capital letter in a</td>
<td>• Supporting ideas</td>
</tr>
<tr>
<td>sentence</td>
<td>• Conclusion</td>
</tr>
<tr>
<td>• missing question mark</td>
<td>• Introductory material</td>
</tr>
<tr>
<td>• missing final punctuation</td>
<td>• Other</td>
</tr>
<tr>
<td>• missing apostrophe</td>
<td>• Transitional words and phrases</td>
</tr>
<tr>
<td>• missing comma</td>
<td>• Repetition of ideas</td>
</tr>
<tr>
<td>• hyphen error</td>
<td>• Topic relationships and technical</td>
</tr>
<tr>
<td>• fused words</td>
<td>quality</td>
</tr>
<tr>
<td>• compound words</td>
<td></td>
</tr>
<tr>
<td>• duplicates</td>
<td></td>
</tr>
</tbody>
</table>

Organization and Development (These items are used to help determine the holistic score, but a separate score for organization and development is not assigned. An essay may not receive a holistic score if Criterion is unable to identify these items.)
Appendix B *Criterion Scoring Rubric* 

**Score of 6:** essay presents a cogent, well-articulated analysis of the complexities of the issue and conveys meaning skillfully. A typical essay in this category

- presents an insightful position on the issue
- develops the position with compelling reasons and/or persuasive examples
- sustains a well-focused, well-organized analysis, connecting ideas logically
- expresses ideasfluently and precisely, using effective vocabulary and sentence variety
- demonstrates facility with the conventions (i.e., grammar, usage, and mechanics) of standard written English but may have minor errors

**Score of 5:** essay presents a generally thoughtful, well-developed analysis of the complexities of the issue and conveys meaning clearly. A typical essay in this category

- presents a well-considered position on the issue
- develops the position with logically sound reasons and/or well-chosen examples
- is focused and generally well organized, connecting ideas appropriately
- expresses ideas clearly and well, using appropriate vocabulary and sentence variety
- demonstrates facility with the conventions of standard written English but may have minor errors

**Score of 4:** essay presents a competent analysis of the issue and conveys meaning adequately.

- A typical essay in this category
- presents a clear position on the issue
- develops the position on the issue with relevant reasons and/or examples
- is adequately focused and organized
- expresses ideas with reasonable clarity
- generally demonstrates control of the conventions of standard written English but may have some errors

**Score of 3:** essay demonstrates some competence in its analysis of the issue and in conveying meaning but is obviously flawed. A typical essay in this category exhibits ONE OR MORE of the following characteristics:

- is vague or limited in presenting or developing a position on the issue
- is weak in the use of relevant reasons or examples
- is poorly focused and/or poorly organized
- has problems in language and sentence structure that result in a lack of clarity contains occasional major errors or frequent minor errors in grammar, usage, or mechanics that can interfere with meaning
Score of 2: essay demonstrates serious weaknesses in analytical writing.

- A typical essay in this category exhibits ONE OR MORE of the following characteristics:
  - is unclear or seriously limited in presenting or developing a position on the issue
  - provides few, if any, relevant reasons or examples is unfocused and/or disorganized
  - has serious problems in the use of language and sentence structure that frequently interfere with meaning
  - contains serious errors in grammar, usage, or mechanics that frequently obscure meaning

Score of 1: essay demonstrates fundamental deficiencies in analytical writing. A typical essay in this category exhibits ONE OR MORE of the following characteristics:

- provides little or no evidence of the ability to understand and analyze the issue
- provides little or no evidence of the ability to develop an organized response
- has severe problems in language and sentence structure that persistently interfere with meaning
- contains pervasive errors in grammar, usage, or mechanics that result in incoherence

---

i 2010. ETS Criterion webpage. http://www.ets.org/portal/site/ets/menuitem.435c0b5cc7bd0ae7015d9510c3921509/?vgnextoid=b47d253b164f4010VgnVCM10000022f95190RCRD (Accessed 03/17/11).


iii Ibid.


v Ibid.

vi 2010. ETS Criterion webpage. http://www.ets.org/portal/site/ets/menuitem.435c0b5cc7bd0ae7015d9510c3921509/?vgnextoid=b47d253b164f4010VgnVCM10000022f95190RCRD (Accessed 03/17/11).


Activating Theory in the Introductory Classroom: Erving Goffman Visits Wisteria Lane

Lisa A. Melander\textsuperscript{1a} and Susan L. Wortmann\textsuperscript{b}
\textsuperscript{a}Kansas State University, Manhattan, KS 66506
\textsuperscript{b}Nebraska Wesleyan University, Lincoln, NE 68504

Abstract

Instructors of large, general education lecture courses face a number of student engagement and learning challenges. In this article, we develop and assess an interactive lecture that introduces a theoretical perspective and three related concepts to two introductory sociology general education classrooms \((n = 433)\). This interactive lecture consists of narrated illustrative clips from the popular culture television series Desperate Housewives, a series of mini-lectures to accompany PowerPoint slides, gapped lecture note handouts, and student interaction. We assess student comprehension at three different points in time: in a pre-survey, in the class period following the interactive lecture, and on the midterm exam two weeks later. Students reacted positively to the exercise and showed an improvement on their scores at both assessment periods.

Keywords: Large interactive lecture, theory, popular culture.

Instructors of large general education college courses face a number of basic challenges. Their classes are often filled with first year students who may find this typically impersonal and standardized instructional format disorienting and overwhelming (Baker, 1976; Bassis & Allen, 1976). These courses often suffer lower attendance, interaction, interest, and engagement (Borden & Burton, 1999). Instructors note that it is difficult to connect with students (Armstrong, 2008; Geske, 1992) and elicit their feedback (Bridges & Desmond, 2000; Wilson & Tauxe, 1986). In these settings, lecture is still the most commonly used format (Mulryan-Kyne, 2010). Lectures help instructors organize materials, provide context for reading materials, and demonstrate enthusiasm about course materials (Cooper & Robinson, 2000); however, they are also limited in their ability to maintain students’ attention (Barak, Lipson, & Lerman, 2006) and to appeal to multiple learning styles (Cooper & Robinson, 2000). Social science instructors also encounter an additional challenge: effectively teaching students to understand, apply, and retain several competing theoretical frameworks such as functionalism, conflict, and symbolic interactionism so they may understand disciplinary knowledge in an organized manner. In general, theory is a difficult subject to teach to students (Holtzman, 2005); and in our experience with previous classes, students became bored and inattentive when we presented these concepts in a typical lecture format. They also showed low comprehension and retention of

\textsuperscript{1} Corresponding author's email: lmeland@k-state.edu
these theories as evidenced by their inability to answer basic recall and application questions on examinations.

We wanted to address these challenges by providing a memorable and engaging educational classroom experience that would allow students to understand, apply, and retain the dramaturgical approach of Erving Goffman which corresponds to the symbolic interactionist theory. To those ends, we developed, piloted, and evaluated an interactive lecture featuring popular culture clips from *Desperate Housewives* in two sections of introductory sociology. In this article, we detail these steps which could easily be adapted for other social science courses in disciplines such as communication studies, women’s studies, psychology, and criminal justice.

**Engagement and Learning Strategies in the Classroom and the Interactive Lecture**

Research suggests that students who are actively engaged in a variety of ways during their college experience increase and even accelerate their learning (National Survey of Student Engagement, 2009). Many scholars have called for active engagement techniques in the classroom to facilitate student engagement and learning (Carpenter, 2006; Chickering & Gamson, 1991; Clark, 2008; Rubin & Hebert, 1998). The extent to which they advocate that students take control of their learning could be placed on a continuum from completely self-guided discovery where students actively construct knowledge on their own through “hands on and free activities” to a more structured process where they are encouraged to select, organize, and integrate knowledge (Mayer, 2004). Techniques for large lecture classes typically fall on the latter end of this continuum due to space and time constraints.

A number of techniques and technologies have been suggested to facilitate engagement and learning in varying class sizes: clickers with constructivist pedagogy (Judson & Sawada, 2002), flip video camcorders (Gao & Hargis, 2010), online blogging (Pimpare & Fast, 2008), music (Albers & Bach, 2003), crossword puzzles (Davis, Shepherd, & Zwiefelhofer, 2009), stories (Levy & Merenstein, 2005), and media clips from soap operas (Hood-Williams, 1986), reality shows (Berry, 2008; Misra, 2000), and cartoons (Scanlan & Feinberg, 2000). In addition, the ubiquitous feature of most large lecture halls, PowerPoint, while critiqued by some for passivity, has been lauded by others as a promising technology of learning (Clark, 2008; D’Arcy, Eastburn, & Bruce, 2009; Pippert & Moore, 1999). For instance, Clark (2008), citing Gilroy (1998), notes that students are socialized to accept the presence of screens and that technology such as PowerPoint can serve to “revitalize” lectures, by facilitating active learning and adding variety to the classroom. In short, previous research has shown that students respond positively when technology is used to enhance traditional lecture materials (Day & Kumar, 2010; Pimpare & Fast, 2008). Furthermore, Marsh and Sink (2010) indicate that students who receive gapped lecture notes report that they more efficiently encode lecture information. Gapped lecture notes are lecture-related handouts that provide a portion of lecture information such as an outline of the topic, headings, key concepts, names and dates as well as space for students to take notes. Together, these studies provide evidence that additional visual
or aural materials combined with a contemporary twist can facilitate student interest by allowing them to see course materials in a different light. Many also highlight the importance of using classroom assessment to determine if and how students learn from interactive formats.

Informed by these approaches, we combined techniques to construct what we call an “interactive lecture” specifically designed to engage students in a large class in a structured manner while appealing to multiple learning styles (Felder, 1996). Our format consisted of a series of three mini-lectures intended to introduce three concepts, each of which were followed with visual materials via PowerPoint featuring applied examples from Desperate Housewives to vividly illustrate our concepts with humor. We also provided gapped lecture notes and asked students to engage in two think-pair-share activities (Lyman, 1981) in which students individually reflect on a prompted topic and then discuss their thoughts with a classmate seated next to them. Volunteers are then invited to share their responses with the larger class. We assessed our project by asking students to answer survey questions immediately before and after our presentation, and a follow-up clicker question during the next class period. They were also tested on concept retention and application on their midterm two weeks later. We discuss the format of our interactive lecture featuring three segments (Performances At the Grocery Store, Region Management at Home, Hiding Stigma), the results of our student responses, the challenges and benefits of this presentation, and suggestions for classroom incorporation.

**Symbolic interaction and Desperate Housewives: Perfect Together**

Symbolic interaction is a primary theoretical framework that focuses on micro-level interaction. In contrast to functionalist and conflict theories which emphasize large scale societal influences on human behavior, this theory allows students to appreciate their own active participation in creating the world around them through social interaction. Dramaturgy, a related approach developed by Goffman, enlists theatrical metaphors to illustrate the ways in which people act in their daily lives (Edgley, 2003). In essence, dramaturgy reveals that the world is a theater where individual actors play multiple roles that create and express their identity. Goffman believed that while enacting these roles, people simultaneously construct a fragile sense of self (Brown, 2005). As a result, people must engage in a number of tactics (e.g., impression control) to preserve and perform the desired self. A dramaturgical lens sheds light on this performance and can be applied to any human communication, even those based upon fictitious events. Scholars have used it to illuminate a number of disparate social interactions such as communication of power in social movements (Benford & Hunt, 1992), the stigma associated with Mardi Gras (Redmon, 2003), and political candidates’ presentation of self (Brown, 2005).

The primetime soap opera, Desperate Housewives, consistently portrays mundane events ripe for dramaturgical analysis: grocery shopping, home life, and neighbor interactions. Pedagogically, we enlisted it for three reasons. First, the larger than life characters and their humorously dramatic situations clearly illustrated the three dramaturgical concepts that we wished students to understand (i.e., performances, region management, and stigma). Second, as a show that averages over 12 million viewers (Nielsen Television, 2011),
its title and characters are part of a shared popular cultural landscape. Third, viewers do not have to be familiar with the plotlines to understand the concepts as Desperate Housewives narrator Mary Alice provides a running background commentary that contextualizes scenes and character motivations. Indeed, her narratives display some of the same bald observations that Goffman made in many of his works, almost as if he was visiting Wisteria Lane.

Methods

Bringing Desperate Housewives to the Masses

We presented our interactive lecture in two large sections of an introductory sociology class at a mid-sized Midwestern university. Students were primarily white, first year undergraduates. They met for 50 minutes twice a week in a lecture hall which held approximately two hundred and fifty students. The IRB approved this study (IRB# 2008028739EP).

Before class, we selected and asked students to read two excerpts from Goffman’s Stigma (1963) and Presentation of Self (1959) that would allow them to encounter the theoretical terminology. We pre-selected and embedded into PowerPoint three short Desperate Housewives clips from Season One that corresponded with each of our key concepts. These were integral to students understanding dramaturgy in particular and symbolic interaction in general. Table 1 describes the clips we used in our interactive lecture.

Table 1. Key Concepts as Illustrated by Desperate Housewives

<table>
<thead>
<tr>
<th>Dramaturgical Concept</th>
<th>Desperate Housewives Episode</th>
<th>Television Clip Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performances</td>
<td>&quot;Pilot&quot;</td>
<td>Lynette lied to a former colleague about enjoying being a full-time mom.</td>
</tr>
<tr>
<td>Region management</td>
<td>&quot;Pilot&quot;</td>
<td>Bree did not show her husband how devastated she was by the news that he wanted a divorce.</td>
</tr>
<tr>
<td>Stigma</td>
<td>&quot;Ah, but underneath&quot;</td>
<td>Gabrielle engaged in stigmatizing behavior when she had an affair with her teenaged gardener.</td>
</tr>
</tbody>
</table>

At the beginning of class, we distributed a brief pre-survey and provided students a gapped handout with summary slides of the dramaturgy concepts. In what follows, we very briefly sketch how we verbally introduced the concepts, provided contextual information before each clip, and briefly reinforced the information through reiteration and discussion.

Performances at the Grocery Store

According to Goffman, it is fundamental to understand that we are all performers. Thus, performance was the first concept that we addressed in our opening mini-lecture, on Po-
werPoint, and in their handouts. As performers, each of us acts to present or maintain an impression of who we are. Because we often feel pressured to perform according to social norms and values, we can become “cynical performers”—people who perform in line with societal standards that they do not necessarily hold (Goffman, 1959). To illustrate these ideas, we showed Lynette Scavo’s performance at the grocery store. Lynette, a formerly successful advertising executive who has recently opted to be a stay-at-home mom, has found that domestic life is much more stressful than the workplace. In our video excerpt, a harried Lynette chases her four unruly children around the grocery store. Suddenly, she bumps into a former co-worker (Cherry, 2004c):

**Former female co-worker:** “So, how’s domestic life? Don’t you just love being a mom?”

**Narrator:** “And there it was. The question that Lynette always dreaded.”

**Lynette:** “Well, to be honest…”

**Narrator:** “For those who asked it, only one answer was acceptable. So Lynette responded as she always did. She lied.”

**Lynette:** “It’s the best job I’ve ever had.”

After the clip, we emphasized that Lynette’s grocery store conversation and the facade of perfect motherhood that she projected illustrated Goffman’s idea of a cynical performance.

**Region Management at Home**

We next wanted to show students how performances can vary by locations, identified by Goffman (1959) as “regions.” Depending on the desired image, people separate their activities into the appropriate front, back, and outside regions. The term “region management” refers to this process. Again, we briefly defined region management using PowerPoint. Then, we segued to a clip of an exceptionally adept region manager, Bree Van De Camp, who routinely engages in region management by projecting a consistent image of the perfect wife and mother. In this scene, when her husband Rex announces that he wants a divorce, Bree retreats alone to the bathroom to cry. After a few minutes, she returns perfectly composed. The narrator then says, “Bree sobbed quietly in the restroom for five minutes. But her husband never knew because when Bree finally emerged, she was perfect” (Cherry, 2004c).

After the clip, we asked students to participate in a think-pair-share activity (Lyman, 1981) by first having them think of an example of region management in their own lives. They then discussed their answers with the person next to them. Finally, we asked for volunteers to share their answers with the class. Students seemed to enjoy the activity as they provided appropriate and humorous examples that provoked audience laughter.

**Hiding Stigma**

Lastly, we wanted students to understand that although people generally attempt to present a consistent self, they sometimes are unable to maintain a desired performance. One
threat to maintaining a successful performance is stigma, which refers to any personal or social attribute that is deeply discrediting (Goffman, 1963). We explained that Goffman was very interested in these presentation failures and the anxiety and anguish that they may produce. We cited the opening example from Stigma (1963) from the students’ assigned reading which details a number of ways stigma alters interaction. It opens with a “desperate” letter from an advice column seeker who cannot have normal social relations because she lacks a nose. The woman signs her letter with only three words: “Sincerely Yours, Desperate.” As such, this letter demonstrates how physical attributes can negatively impact social interaction.

Thirty-something former model Gabrielle Solis does not have such identifying physical blemishes. Instead, her adulterous affair with her underage gardener could cause her to be stigmatized in public and in private. In our clip, Gabrielle and a scantily clad teenage gardener are in her bedroom when her husband Carlos unexpectedly returns home. Gabrielle quickly throws the teenager and his clothing out the window. According to the narrator, “What her husband couldn’t see couldn’t hurt her” (Cherry, 2004a). Thus, for the moment Gabrielle has successfully hidden her stigmatizing behavior. After this clip, we again asked students to share an example of stigma (either physical or behavioral) with the person sitting next to them. Again, the lecture hall was buzzing with student conversation. Several students then volunteered good examples of stigma.

At the end of class, students completed a post-survey of their perceptions of the interactive lecture. During the class period immediately after the presentation, students responded to a follow-up clicker question on these materials. We present the results of our pilot interactive lecture assessment in the following section.

**Assessment**

Our assessment combined pre- and post-survey results from both classrooms. Both contained measures of how thoroughly students read the assigned materials, whether they understood and retained the key concepts, and if they liked the theoretical excerpts and presentation. Two of the concept questions were also included on the midterm exam. We compared the midterm responses to the pre-survey responses to measure concept retention.

Because their survey responses were anonymous, we were unable to restrict our clicker responses and midterm examination scores to include only those who attended the interactive lecture. Furthermore, class attendance was not mandatory, and the number of respondents fluctuated between the assessment periods. As such, 433 students participated in the interactive lecture and pre- and post-surveys, 377 responded to a follow up clicker question in the next class, and 471 took the midterm exam. All of the students consented to have their responses used for research (n = 433).
**Student Preparation and Interest**

In a large lecture setting, instructors often wonder whether students read the assigned materials. We were pleasantly surprised by the pre-survey results that showed a majority of the students (53.4%) reported reading the entire articles before class. Only about 7% \((n = 30)\), however, felt confident in their ability to explain any two main ideas from the article. Their confidence levels sharply increased after the interactive lecture: 62.5% \((n = 271)\) indicated on the post-survey that they were very confident or confident in their ability to explain two main concepts, and 34.2% \((n = 148)\) noted that they felt fairly confident. We were encouraged by these results, especially because previous sections of students struggled with theoretical concepts.

In our experience, students struggle with theories because they seem inaccessible and removed from their lives. At the beginning of the lecture, we asked for a show of hands of those who liked learning theories. No one raised their hand in the first class, and only a handful from the second responded affirmatively. We directly asked why they were not interested in theory and several volunteered “it’s boring” and “it’s too hard” while others nodded their heads in agreement. We said we hoped to change their opinions. We were not surprised that the pre-survey revealed only 20.6% \((n = 88)\) reported interest in the assigned materials while most (43.1%; \(n = 184\)) reported feeling “indifferent” to them.

Table 2. Comparison of pre-survey and midterm exam scores.

<table>
<thead>
<tr>
<th>Goffman Concept</th>
<th>Percentage of Class Correct on Pre-Survey</th>
<th>Percentage of Class Correct on Midterm</th>
<th>Change Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performances</td>
<td>63.0% ((n = 271))</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Region Management</td>
<td>16.80% ((n = 72))</td>
<td>47.60% ((n = 224))</td>
<td>30.80%</td>
</tr>
<tr>
<td>Stigma</td>
<td>54.20% ((n = 232))</td>
<td>75.50% ((n = 355))</td>
<td>21.30%</td>
</tr>
</tbody>
</table>

*Note: Students were only asked a question about performances on the pre-survey.*

**Student Comprehension and Retention**

The scores on three pre-survey concept items indicated that the students understood or retained little from their own reading (please see Table 2). Only 16.8% \((n = 72)\) could correctly identify an example of region management. More students selected the correct answers for question about the best example of stigma and performances (54.2% \((n = 232)\) and 63% \((n = 271)\), respectively). In the class period following the interactive lecture, students responded to one clicker question that asked them to apply the three concepts previously covered to a different scenario. Approximately 80% \((n = 301)\) selected the correct response, although probably not all of the clicker respondents were present for the interactive lecture.
As an indication of their longer-term retention, we included two of the pre-survey concept items on the mid-term exam two weeks later. Even though different numbers of students participated in the phases of this project, the percentage of students who selected correct answers greatly increased on the midterm. For example, 75.5% \((n = 355)\) selected the correct response for the stigma question on the mid-term exam compared to 54.2% \((n = 232)\) on the pre-survey. Similarly, the percentage with correct answers on region management \((47.6\%, n = 224)\) increased. It is possible that a higher percentage correctly answered the “stigma” question because this term is commonly used outside of sociology.

**Student Engagement**

Overall, the students appeared to be very attentive during class and reacted positively to the interactive lecture. In the post-survey, 84% \((n = 365)\) indicated that the presentation was “interesting” and 83% \((n = 358)\) recommended that we use the presentation in future classes. Indeed, several approached us after each class, informing us that they enjoyed and had learned from the presentation. One emailed us with suggestions for additional scenes. Nearly two months later on end of semester course evaluations, a handful of students from both classes wrote positive comments specifically about the interactive lecture. Because students seemed engaged during this lecture, had positive comments about the activity, and were able to retain this material for the midterm exam, we plan to use this material and our interactive method in future classes.

**Discussion**

As Holtzman (2005) observes, students are often resistant to theory. Instructors of larger general education courses in particular must be innovative when presenting difficult but foundational concepts, and students often respond well to an entertainment factor in the classroom (Delucchi & Pelowski, 2000). For this interactive lecture, we deliberately selected materials and combined technologies and strategies to introduce Goffman’s concepts not only to engage and involve our students in understanding a key theoretical framework, but also to provide memorable examples that might help them to retain this information. In this paper, we outlined our approach that combines a series of mini-lectures around three concepts enhanced by popular culture examples, handouts, and peer discussion. We also evaluated the interactive lecture at different points in the semester through paper and clicker formats. Although these additional steps were time consuming, evaluation of the methods helped us to assess whether or not we were moving towards our educational goals. Our pilot results suggest that this interactive lecture was well received and that our method may encourage comprehension and retention.

We would be remiss not to briefly note technological, methodological, and content concerns associated with this activity. Technologically, it is challenging and time consuming to locate and edit television examples. Websites such as abc.com, hulu.com, youtube.com, or netflix.com can be enlisted to locate appropriate video materials. Methodologically, it is also important to have the same respondents participate in each assessment phase. In the future, we would like to match respondents at the different assessment periods to determine more definitively whether bringing Goffman to Wisteria Lane has a
positive impact on student retention of these materials. Additionally, we may include a comparison group and concept items on the final exam to assess longer term retention.

Finally, while *Desperate Housewives* has inspired academic analysis (McCabe & Akass, 2006), its content has also been the subject of critique because of its stereotypical gender images (Pozner & Seigel, 2005). Popular cultural images can be read in multiple ways—as sources of oppression or of subversion and resistance. As gender scholars, we are aware of our responsibility to not only discourage stereotyping but also to enable students to critically view cultural images. We addressed these pitfalls in three ways in our classroom. First, we acknowledged at the outset of the lecture that *Desperate Housewives* presents stereotyped gender images. Second, we encouraged students several times throughout the semester to critically examine and discuss media images and stereotypes. At those times, we used the characters as an example of stereotypical images. Third, we offered an extra credit assignment examining gendered media images using the DVD *Dreamworlds III* (Jhally, 2007). In essence, bringing *Desperate Housewives* to the classroom actually provided further opportunities to analyze gendered images.

There are also important strengths associated with introducing theory and other academic topics via popular culture exemplars such as *Desperate Housewives* in other classes. For example, criminal justice students can learn more about electronic monitoring devices as well as the personal and familial challenges associated with house arrest by viewing a *Desperate Housewives* scene featuring Carlos and Gabrielle on this topic (Cherry, 2004b). Like Misra (2000) and Scanlan and Feinberg (2000), we believe that using television shows is an effective approach to pique student interest. Students often feel alienated from instructors in large classrooms because the instructor seems to be a remote authority figure (Geske, 1992). Our use of a popular show may have supplied an important link to our students as they were able to see that we watched and enjoyed common television programs. Additionally, the *Desperate Housewives* scenes provided an interesting and relevant popular culture application of theoretical concepts and may have assisted with long-term concept retention and piqued their interests in theoretical topics. The format used in this article also allows instructors to assess student performance and comprehension at different time periods using multiple methods. Finally, although *Desperate Housewives* may present stereotypical situations and images, bringing it into the classroom allowed us to encourage students to think critically about these and similar popular culture images and messages at several later points in the semester.

References


Group Work: How to Use Groups Effectively

Alison Burke1
Southern Oregon University, Ashland, OR 97520

Abstract

Many students cringe and groan when told that they will need to work in a group. However, group work has been found to be good for students and good for teachers. Employers want college graduates to have developed teamwork skills. Additionally, students who participate in collaborative learning get better grades, are more satisfied with their education, and are more likely to remain in college. This paper will discuss the use of group work in higher education.

Keywords: Group work, collaborative learning, higher education pedagogy.

Teaching and learning in higher education are changing. Active learning has become an important focus in this time of pedagogical change. While the term encompasses a broad array of practices, collaborative learning, or small group work, remains an important element of active learning theory and practice. Research suggests that students learn best when they are actively involved in the process (Davis, 1993). According to Wasley (2006), “Students who participate in collaborative learning and educational activities outside the classroom and who interact more with faculty members get better grades, are more satisfied with their education, and are more likely to remain in college” (p. A39). A collaborative learning environment, as opposed to a passive learning environment, helps students learn more actively and effectively (Murphy, Mahoney, Chen, Mendoza-Diaz & Yang, 2005). Additionally, research also shows that employers want college graduates to possess the ability to work in groups and have developed suitable teamwork skills (Blowers, 2000).

This paper is designed to offer suggestions on how to use small groups in order to facilitate learning and instructional diversity in face to face classes. It will begin with an overview of the advantages and disadvantages of working in a group and then discuss some of the methods available to enhance group learning and communication. It will also provide some suggestions for evaluating and assessing group work. While many people detest the mere suggestion of group work, it can be an effective tool if used appropriately.

Grouphate: What’s there to love?

Many people cringe and groan when told that they will need to work in a group. This phenomenon is called “grouphate.” Grouphate has been referred to as the dread and re-
pulsion that many people feel about working in groups or teams (Sorenson, 1981). However, these feelings diminish among group members who have received proper instruction about working in groups. One way to overcome group hate is to form realistic expectations of group work. According to Beebe and Masterson (2003), there are advantages and disadvantages to working in a group. By understanding the benefits and potential pitfalls, a group can capitalize on the virtues of group work and minimize the obstacles that hinder success.

Advantages: There are six advantages to working in a group:

1. Groups have more information than a single individual. Groups have a greater well of resources to tap and more information available because of the variety of backgrounds and experiences.
2. Groups stimulate creativity. In regard to problem solving, the old adage can be applied that “two heads are better than one.”
3. People remember group discussions better. Group learning fosters learning and comprehension. Students working in small groups have a tendency to learn more of what is taught and retain it longer than when the same material is presented in other instructional formats (Barkley, Cross & Major, 2005; Davis, 1993).
4. Decisions that students help make yield greater satisfaction. Research suggests that students who are engaged in group problem solving are more committed to the solution and are better satisfied with their participation in the group than those who were not involved.
5. Students gain a better understanding of themselves. Group work allows people to gain a more accurate picture of how others see them. The feedback that they receive may help them better evaluate their interpersonal behavior.
6. Team work is highly valued by employers. Well developed interpersonal skills were listed by employers among the top 10 skills sought after in university graduates (Graduate Outlook Survey, 2010).

Disadvantages: Although working in groups has its advantages, there are also times when problems arise. Beebe and Masterson (2003) list four disadvantages.

1. There may be pressure from the group to conform to the majority opinion. Most people do not like conflict and attempt to avoid it when possible. By readily acquiescing to the majority opinion, the individual may agree to a bad solution just to avoid conflict.
2. An individual may dominate the discussion. This leads to members not gaining satisfaction from the group because they feel too alienated in the decision making process.
3. Some members may rely too heavily on others to do the work. This is one of the most salient problems that face groups. Some members do not pitch in and help and do not adequately contribute to the group (Freeman & Greenacre, 2011). One solution to this problem is to make every group member aware of the goals and objectives of the group and assign specific tasks or responsibilities to each member.
4. It takes more time to work in a group than to work alone. It takes longer to accomplish tasks when working with others. However, the time spent taking and analyzing problems usually results in better solutions.

Overall, effective student participation in group work is an important learning outcome for higher education courses (Elgort, Smith & Toland, 2008). Although many students feel as though they can accomplish assignments better by themselves rather than in a group, instructors find that group work helps the students apply knowledge (Elgort, Smith & Toland, 2008). However, merely assigning a group does not itself create critical thinking outcomes. Therefore, the instructor must be cognizant of how best to facilitate effective collaborative learning environments.

There are four stages of group work. First, the instructor must decide that he/she wants to incorporate group work into the class. The group work should be designed into the syllabus. The second stage involves teaching the students to work in a group. Instructors cannot assume that students know how to work together, structure time, and delegate tasks. The instructor must be able to teach the students how to work proactively in groups. This leads to the third stage, which involves monitoring the groups. The last stage, and the most important to the students, is the assessment of the group. The instructor must develop a concrete rubric for grading the students.

**Getting Started**

The best place to start group work (much like anything else) is at the beginning. When developing a course syllabus, the instructor can determine what topics and theme lend themselves to group work. This is the time that instructors can think about how they will form their groups, help negotiate the group process, and decide how to evaluate the final product.

Johnson, Johnson and Smith (1991) suggest that group tasks should be integral to the course objectives. This means that the group work should complement the learning objectives outlined in the syllabus. If one of the learning objectives is to promote critical thinking skills or writing enhancement, then the group work should support these areas.

**Group Size**

The dynamics of group size is an important component of group work. A small group is often considered to consist of three or more people (Beebe & Masterson, 2003). Groups of two are called dyads and are not encouraged for group work because there are not a sufficient number of individuals to generate creativity and a diversity of ideas (Csernica et al., 2002). In general, it is suggested that groups of four or five members tend to work best (Davis, 1993). However, Csernica et al. (2002) suggests that three or four members are more appropriate. Larger groups decrease each member’s opportunity to participate and often results in some members not actively contributing to the group. In situations where there is a shorter amount of time available to complete a group task, such as an in-class collaborative learning exercise, it is suggested that smaller groups are more appro-
priate. The shorter amount of time available, the smaller the group should be (Cooper, 1990; Johnson, Johnson & Smith, 1991).

Group work can be especially beneficial for large classes. Wright and Lawson (2005) found that group work helped students feel that the class was smaller and encouraged them to come to class more often. They felt more invested in the course and in the class material, which promoted active learning in a large class environment.

Assigning a Group

Assigning the members of the group is integral to the success of the group. Some faculty members prefer to randomly assign students to groups. This has the advantage of maximizing heterogeneity of the group (Davis, 1993) and is an effective way of assigning group members in large classrooms. If the class size is small and the instructor is familiar with most of the students, the instructor can select the group members based on known attributes of the class. For example, the instructor can form the groups while taking into account performance levels, academic strengths and weaknesses, ethnicity, and gender (Connery, 1988).

Additionally, some instructors allow the class the self-select their group; however, this has some disadvantages. Self-selected groups often gravitate toward friends and roommates (Csernica et al., 2002). This can result in the students self-segregating and spending more time socializing than working on the group project (Cooper, 1990). Research suggests that groups which are assigned by the instructor tend to perform better than self-selected groups (Felder & Brent, 2001).

Teaching Students

It is difficult for teachers to design and implement group work effectively, and it is difficult for students to foster the group process, especially if they do not have the skills to make effective use of group work. Many students have never worked in a group before or lack the skills to work with others. Instructors cannot assume that students know how to work together, structure time, or delegate tasks. There are several ways that instructors can help.

First, the instructor should make certain that each student understands the assignment. Students should know the purpose of the project, the learning objective, and the skills that need to be developed through group work. Successful group work is easier if the students know how the assignment relates to the course content and what the final product is supposed to be (Davis, 1993).

Second, the instructor needs to reinforce listening skills and the proper methods to give and receive constructive criticism. These skills can be discussed in class and modeled during class activities. Some faculty use various exercises that are geared toward helping students gain skills to work in groups (Fiechtner & Davis, 1992). Small in-class group activities help reinforce cohesion and group unity.
Third, the instructor needs to help the students manage conflict and disagreements. The instructor should avoid breaking up the groups (this will be discussed in more detail later in this paper). When a group is not working well together, the students need to learn how to communicate effectively and establish goals for a successful group (Davis, 1993).

**Monitoring the Group Process**

One method to help groups succeed is to ask each group to devise a plan of action (Davis, 1993). The plan of action involved assigning roles and responsibilities among the group members. Each member should have a role, such as the note taker or the group spokesman. The instructor can review each group’s written plan of action or meet with each group individually and discuss their plan.

Another method to help monitor a group’s progress is to ask them to submit weekly progress reports. These reports (or weekly meeting notes) should outline what the group discussed, who attended the meeting, and the objectives set for the next week. In this manner, the instructor can monitor the group’s activities and progress throughout the semester and assess the level of involvement from each member.

**Group Dissonance**

Groups will not always work well together. Some groups lack motivation, strong leadership, or simply have personality conflicts. Even when it appears that a group is falling apart, it is important to avoid breaking up the group. Not only will the group dynamics of the original group be affected if the members are reassigned, but the addition of members to other groups will disrupt their dynamics as well (Davis, 1993).

One way to help prevent conflict and group members who shirk duties is to keep the group small. It is difficult to be a “loafer” or a “slacker” in a small group (Davis, 1993). Additionally, matching work assignment to skill sets will help separate the “loafers” from the students who are generally struggling (Freeman & Greenacre, 2011). Freeman and Greenacre (2011) suggest that instructors should help the students understand the benefits of working together as a group for the group as a whole, which will help students who are struggling (Freeman & Greenacre, 2011). Furthermore, the group should be encouraged to have assigned roles and responsibilities. It is more difficult to be a slacker if the goals are clearly outlined for each member.

It is necessary to help a group work through disagreements and find resolution. Simply breaking up the group does not encourage the students to work though differences. Freeman and Greenacre (2011) suggest that group interventions should be aimed at the destructive group member, focus on the behavior and not the person, and address the benefits of the group process for the group as a whole. Barkley et al. (2005) recommend designing the coursework in such a way that the success of the individual relies on the success of the group. The instructor should assist the group in creating ways in which to handle unproductive members and foster communication skills.
Evaluation

Evaluating a group is a difficult task and the instructor should have a clear idea of how he/she wants to evaluate the group work. First, the instructor should decide what is being evaluated: the final product, the process, or both. Next, it is necessary to decide who assigns the grade: the students, the instructor, or both. Some faculty members assign each member of a group the same grade, which may promote unhappiness if some members devote more time and effort to the group and get the same grade. Some instructors assign each group member an individual grade, which may or may not foster competition within the group and may undermine the group solidarity (Davis, 1993). If the group is graded as a whole, it is suggested that the project or presentation should not count for more than a small percentage of the student’s final grade (Cooper, 1990; Johnson, Johnson & Smith, 1991).

If the aspect of process is going to be evaluated, it is important to give the students an opportunity to assess the effectiveness of their group. At the end of the process, they should be able to list their contributions, their group member’s contributions, and the process as a whole. They should be able to identify the aspects that worked and the aspects that did not work. The student’s group assessment allows the instructor to evaluate the group process and apply the most effective methods to future group projects.

In the aspect of assessment, it is vital that the students know and understand how they will be evaluated. One method used to convey this information is with a structured grading rubric. A rubric is a scoring tool which lists the criteria by which a paper or presentation will be graded. The rubric lists, not only the criteria by which the work is judged, but also the student’s mastery of the material (Finson & Ormsbee, 1998). Stevens and Levi (2005) advocate the use of rubrics because they: convey expectations to the students, help students focus their efforts, improve student achievement, reduce grading time for the instructor and improve the effectiveness of feedback.

If the instructor is interested in assessing the group process and final product, two separate rubrics need to be created. For the process, the evaluation criteria should represent the learning objectives for class and for the group. Process evaluation might include: attendance and participation in meetings, time management skills, active listening, evidence of cooperative behavior, and professionalism and engagement with the task.

For evaluating the effectiveness of the product, a more concrete grading rubric might be necessary. The criteria can be outlined based on content, structure, organization, accuracy, thoroughness, and general mechanics. Rubrics can be helpful for both students and instructors; they outline expectations and allow instructors to assign grades on a more objective basis. Rubrics provide detailed breakdowns of points that are awarded for each criterion and how those points are awarded. Additionally, rubrics are useful beyond grading; they also help students conceptualize the assignment (McKeown, 2011).

The instructor knows what the end product should look like, and it is his or her responsibility to effectively convey that expectation to the students (Finson & Ormsbee, 1998).
Herman, Aschbacher, and Winters (1992) suggest that a good rubric consist of clear and logical categories that explain what the instructor is assessing and the point value for each response. This practice ensures that both the student and the teacher are satisfied with the end result and makes the group process a more objective learning experience. Rubrics are very useful for evaluating group work and increase the chances of student success (Finson & Ormsbee, 1998). McKeown (2011) asserts that rubrics should be locally relevant and culturally appropriate. In this way, they cannot be a “one size fits all” approach and must be modified to fit different classes, disciplinary perspectives, and learning objectives.

Conclusion

There is a well of information about group work and the benefits of collaborative learning. When students spend time meeting in groups, they are able to achieve a deeper learning themes covered in class as well as develop skills, such as writing and communication (Light 2001). Wright and Lawson (2005) found that the bridging of in- and outside-class work encourages students to spend more time preparing for class, and having conversations with team members outside of regular class time.

One of the most recent trends in collaborative learning appears to be the use of technology to advance group work. Elgort, Smith and Toland (2008) utilized the use of wikis in a class and found that most wikis encouraged student participation and group involvement, but did not counteract student preference to work alone rather than in a group. Additionally, many instructors use online discussion forums in their face to face classes to foster participation and engagement (Wright & Lawson, 2005). This has the advantage of engaging commuter students and students who have jobs in addition to taking classes. Utilizing online or web based might help resistant students engage in and benefit from group work.

Overall, it seems clear that when it comes to group work, the whole is greater than the sum of its parts. While creating, monitoring, and evaluating groups is a recursive process, active learning techniques are beneficial for students. Supplementing lectures with group work helps students feel engaged and subsequently learn more (Payne, Mott-Turner, Smith & Sumter, 2004). Group work helps students develop teamwork skills and social interactions as well as learning about various backgrounds, culture, beliefs, and attitudes (Payne et al., 2004). Group work does not have to yield “group hate” and as long as the instructor is properly prepared to introduce and facilitate group involvement and participation, group work can produce very positive and lasting results.

References


The Absent Graduate Student: An A-B-A Single-Subject Experiment of Online Discussion Participation

Thomas D. Cox
University of Houston-Victoria, Victoria, TX 77901

Abstract

It can reasonably be said that the discussion board is the “heart” of an online course. The discussion board can and should help the learner to make meaning of the content or topic and to find relevance to their lives. When this “engagement” does not occur, the individual does not benefit from the discussion, and it becomes merely “busy work” and no deep learning takes place. The primary purpose of this study was to investigate an individual master’s student who was not participating in the course discussion and apply a single-subject experimental design (A-B-A) in an effort to increase the student’s participation in the course. The method was to observe the baseline behavior, apply the treatment, withdraw the treatment, and then re-assess the baseline behavior. The results of the experiment were astounding to the researcher. Not only was the experiment successful with the non-performing student, but other students in the course also dramatically increased their level of participation. The results of this study can serve as a best practice to remedy the sometimes disembodied nature of online learning.

Keywords: Individual discussion participation, engaging best practices, modeling discussion, teaching strategies, effective engagement.

If the discussion board is the “heart” of the online course, then the professor is the lifeblood that nourishes the body—the student body. Many professors use course management systems to deliver course content, but some are still confused by the use of threaded discussion (Palloff & Pratt, 2007). Discussion about topics learned is vital for a thorough acquisition of knowledge, whether the class is a hybrid or is conducted entirely in cyber-space (Bender, 2003).

Literature Review

In Bloom’s (1956) taxonomy, “knowledge” is the lower foundational level of a sequence of progressive contextualization of material. For purposes of this introduction, the discussion board is where the student can place material given by the professor into a context, make meaning of the material, and then progress upward through Bloom’s taxonomy toward comprehension, application, analysis, synthesis, and evaluation (Figure 1). With this understanding of the importance of the discussion to online learning, professors are

1 Corresponding author’s email: CoxT@uhv.edu
charged with finding ways to encourage and motivate students to participate in the discussion board.

<table>
<thead>
<tr>
<th>Higher Order Thinking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluation</td>
</tr>
<tr>
<td>Synthesis</td>
</tr>
<tr>
<td>Analysis</td>
</tr>
<tr>
<td>Application</td>
</tr>
<tr>
<td>Comprehension</td>
</tr>
<tr>
<td>Knowledge</td>
</tr>
<tr>
<td>Lower Order Thinking</td>
</tr>
</tbody>
</table>

**Figure 1. Adapted from Bloom’s Taxonomy (1956)**

A professor who proactively encourages and motivates students to actively engage in the online discussion thread has intentionally or unintentionally taken on a certain role. McKeachie (1978) describes six roles for the campus teacher: facilitator, expert, formal authority, socializing agent, ego ideal, and person. Bender (2003) adds, “All of which can be applied to online teaching” (p. 11). These roles can all come into play at different times and in different components of an online course, but the author suggests that for the discussion board, the role of facilitator is the primary role. In this role, says Bender, the facilitator enhances student learning by encouraging active participation in discussion. See Table 1 below.

Robles and Braathen (2002) as cited in Fjeltstul, Tesone, and Bougac (2008) in an article in *The Journal of Online Learning and Teaching* (JOLT) place additional importance on the threaded discussion. “Discussion in a traditional face to face environment allows for knowledge sharing between and among instructors and students. Similar discussions take place in an online environment; however, such communication is through threaded discussions. Certainly, arguments can be made for benefits and challenges of both environments. Threaded discussions afford students the flexibility of their engagement, time for thorough reflection and thought prior to engagement, and eliminate the often uncomfortable public speaking that reduces and sometimes all together prevents students from active participation in class discussion. Students also have the ability to respond to every question/issue raised in the online learning environment” (p.293).

The instructor’s role in personalizing the online learning experience cannot be understated. Gallien and Oomen-Early (2008) remind us that there is no doubt that instructor feedback is important to enhance student learning, and recent findings have shown that online learners’ levels of satisfaction, performance, and sense of community are related to the interactions they have with their instructors, including the type and frequency of feedback they receive on assignments and course material.
From this, it can be seen that the role of the professor is crucial and important to the discussion as well as the benefits/advantages to students when they actively engage in the discussion board.

**Table 1. Adapted from Bender (2003) and McKeachie (1978).**

<table>
<thead>
<tr>
<th>Role</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facilitator</td>
<td>A facilitator enhances online learning by encouraging active participation in discussion and by helping the student to see education as meaningful and relevant.</td>
</tr>
<tr>
<td>Expert</td>
<td>An expert communicates expertise through lectures and discussions and is able to stimulate students without overwhelming them.</td>
</tr>
<tr>
<td>Formal Authority</td>
<td>A formal authority helps students by establishing boundaries such as acceptable conduct and dates of submission.</td>
</tr>
<tr>
<td>Socializing Agent</td>
<td>A socializing agent who has contacts within the larger academic community, and as such can be helpful to students in providing such things as letters of recommendation and links to research and publication resources.</td>
</tr>
<tr>
<td>Ego Ideal</td>
<td>An ego ideal is charismatic and shows commitment and enthusiasm not only to the subject matter, but also to the students themselves.</td>
</tr>
<tr>
<td>Person</td>
<td>A person who demonstrates compassion and understanding of student needs.</td>
</tr>
</tbody>
</table>

The threaded discussion is a public place for discussion that allows time for reflection. While there is a flow of discussion and it is linear, it is not subject to the tyranny of the ever present "now" of the face-to-face classroom that doesn't allow the participants the benefit of an "instant replay." The discussion forum allows as many replays as a participant wants of what was said. A discussion can be revisited and commented on as long as the forum is open, while in a classroom, often the moment is lost and is difficult to revisit.

With the incorporation of Bloom’s (1956) six hierarchical levels: knowledge, comprehension, application, analysis, synthesis, and evaluation, threaded discussion can be improved. These levels build upon each other as the learner gains knowledge and expertise therefore leading the student to complex understandings and knowledge (Christopher, Thomas, & Tallent-Runnels, 2004). Using complex, higher order questions in the threaded discussion will not only force the student to flex intellectual muscles when re-
sponding, but will also lead the student to more understanding and less recitation (Foote, 2001; Lord & Baviskar, 2007). Depending on the learning objective, using Bloom’s Taxonomy will provide a starting place for the instructor in designing an appropriate level of question.

Instructor response and student response are the key components to the construction of shared knowledge within the discussion forum. The deepest learning is in the writing and "talking" about the content of the course within the community of learners. It is a pedagogically sound practice, based on cognitive learning theories, to design and engage in discussion forums with students. More opportunities are provided for students to become actively involved with the course content to construct their own deeper and lasting learning.

Realizing the crucial role of the discussion thread in an online course, and accepting the role as “facilitator” of that discussion, it then becomes imperative for professors to find ways to involve students in online threaded course discussions. The purpose of this study was to conduct an experiment that was designed to engage an “absent” student in the online threaded course discussion.

Method

In the fall semester of 2009, the researcher was teaching an online graduate course entitled, “The Foundations of Adult Education”. The 15 week course included weekly discussions. The weekly topics were posted seven days prior to the beginning of the discussion. The course “week” was designed to be Tuesday through the next Monday. The students were emailed again at the beginning of the week to inform them of the topic, and remind them to begin their discussion after they had read the material in the text or other articles posted in the course. The students then had seven days in which to discuss the topic.

It was the practice of the professor to e-mail students individually whose participation was not sufficient in the course. It was also the practice of the professor to respond to each student post at least one time by adding information, asking questions, providing support, or redirecting the conversation.

“A-B-A Single Subject Experiment” vs. the “Case-Study” Approach

Single case research comes in two basic varieties, single case experimental designs and case studies. Single-case experiments are used most frequently to study the effects of basic learning processes and to study the effectiveness of behavior modification. A case study is a detailed, descriptive study of a single individual, group or event. The case is described in detail, and conclusions, solutions, or recommendations are offered (Leary, 2010). Since a treatment was applied in this study, the method is said to be experimental, and not simply a “case study”.
The baseline behavior (lack of participation in the discussion) is assessed; the treatment is provided (professor directly engaging the subject), and then withdrawn, and the baseline behavior is reassessed.

Figure 2 is a representation of this type of experimental design where “O” represents and observation or measurement, and “X” represents an exposure to an experimental variable or event, and the effects are to be measured.

<table>
<thead>
<tr>
<th>Baseline A</th>
<th>Treatment B</th>
<th>Baseline A</th>
</tr>
</thead>
<tbody>
<tr>
<td>O-O-O-X-X-X-O-O-O</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 2. Representation of A-B-A Method**

**Limitation of the Study**

This experiment was conducted only to increase the participation of one student in the course. It was not the intention of the study to increase overall student discussion or to imply these methods would work for all students; therefore, general conclusions cannot be drawn about the effectiveness of the treatment. It could also be stated that more participation does not mean better participation, as the experiment was only geared toward increasing participation in the discussion.

**Baseline**

The researcher identified the subject of the study, who will be called “Mary” as having “less than satisfactory” participation in the discussion board. The baseline data was collected for 3 consecutive “sessions” in this case, week one through three. Mary’s baseline behavior was that of established lack of participation; total discussion posts-two.

Table 2 shows a breakdown of Mary’s discussion participation for Week 1 through Week 3. The design involved multiple observations of a single individual. The target behavior of a single individual is established over time and is referred to as a baseline behavior (Creswell, 2009).

It was clear from weeks one through three that teacher intervention was necessary in addition to simply emailing Mary of her performance issues. After three weeks of insufficient participation, the professor emailed Mary and explained that in Week 4, he wanted her input on a few topics and he needed her help to keep the discussion going.

**Treatment**

The treatment was then administered in weeks four, five, and six of the course. The treatment consisted of posting a question in the discussion board on the first day of week
Table 2. Baseline Observations

<table>
<thead>
<tr>
<th>Week 1</th>
<th>Week 2</th>
<th>Week 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Post</td>
<td>1 Post</td>
<td>0 Posts</td>
</tr>
<tr>
<td>2 Sentences</td>
<td>1 Sentence</td>
<td></td>
</tr>
<tr>
<td>No references to readings or material</td>
<td>No references to readings or material</td>
<td></td>
</tr>
<tr>
<td>Only opinion based statements such as “Good point”, “I agree”, etc.</td>
<td>Only opinion based statement, “Very Interesting”.</td>
<td></td>
</tr>
</tbody>
</table>

4 with the subject line “Hello Mary” and subsequently directly addressing Mary throughout the week. This solicited Mary’s direct response. (Note: In an effort not to be cruel or single Mary out, all posts to all students by the professor began with the students’ first names, as to get their attention). The second day of class in week 4, when the professor checked in on the discussion, Mary had responded to the prompt and her response included the answer to the question directly from the text. This was a sign of improvement. The Professor then addressed Mary directly in the discussion by thanking her for a correct and timely response, and asked her what she thought about “John’s” critique of the theory we were studying. The third day the professor logged in and there was a post from Mary, although brief, saying that she thought “John” had done a good job of describing the theory under discussion. We were making progress with getting Mary to participate. In the first week of the treatment (week 4 of the course), Mary had already made a noticeable improvement in her engagement in the discussion. Even though Mary seemed to only respond when the professor called her by name in the discussion, she was in fact being “trained” on how to engage. Week five and six of the course progressed with the professor continuing to engage Mary individually in a no-threatening way, and at the end of week 6, Mary had completely “caught on” and was engaged in the course. Table 3 below describes Mary’s engagement during the treatment. Notice the difference that the treatment made in Mary’s performance from week’s one through three in Table 2 above. It should also be noted here that perhaps the intervention as prescribed to Mary may have well been conducted for all students in Week 1 instead of waiting until week 3 or 4.

As noted previously in Table 1, the professor may take on several roles in order to engage students in active discussion participation. In this case, the professor took on two major roles. First, the professor became the facilitator, or the one who encourages active participation. This involved moving beyond the authoritarian role of simply stating the requirements in the syllabus. Second, in a non-threatening way, the professor took on the role of “person” by demonstrating interest in Mary’s contribution to the course. In essence the teacher modeled the consistent communication that is needed to engage Mary in the discussion and make the course more meaningful to her. As a result, Mary took responsibility and realized that the discussion was a necessary and meaningful part of the course.
Table 3. Treatment Observations

<table>
<thead>
<tr>
<th>Week 4</th>
<th>Week 5</th>
<th>Week 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 Posts</td>
<td>3 Posts</td>
<td>4 Posts</td>
</tr>
<tr>
<td>Post 1 – Two sentences based on terms from the reading.</td>
<td>Post 1 – Three sentences – direct responses to the professor about the reading.</td>
<td>Post 1 – Three sentences in response to the topic.</td>
</tr>
<tr>
<td>Post 2 – One comment about another student’s work.</td>
<td>Post 2 – Asked the professor a question.</td>
<td>Post 2 – Two sentences – asked a classmate to clarify a question.</td>
</tr>
<tr>
<td>Post 3 – One sentence – only opinion based.</td>
<td>Post 3 – One sentence – replied, “Thank you” to the professor’s response.</td>
<td>Post 3 – Posted and opinion on the discussion topic.</td>
</tr>
<tr>
<td></td>
<td>Post 4 – Posted an opinion only statement, “Good point!”</td>
<td>Post 4-</td>
</tr>
</tbody>
</table>
of the course about his concerns. Mary may also have simply needed time to adjust to the beginning of the semester, and getting in the “swing” of things with a new teacher. Also, Mary’s prior experience in classes may have lead her to negate the importance of the discussion board in the course.

The method used may also have other limitations. There is no replication of the “B” condition on Mary. Since there are no replications within the design, there needs to be several A-B-A investigations to support the findings. Also, triangulation, in the form of an individual “needs assessment” or a questionnaire administered to Mary may have revealed the reasons for her lack of participation. For example, it was not known if Mary had even taken an online course prior to this experience.

Results

The results of this A-B-A single-subject experiment showed that the baseline was affected by the treatment. The subject in the study showed noticeable improvement in the level of activity and engagement in the discussion board as evidenced by more posts per week, questions to the professor and classmates, which showed a deeper interest in learning, and posting supportive comments about classmates’ postings. Weeks one through three without teacher active solicitation and engagement resulted in three total posts. Weeks four through six, with the professor’s active engagement of the student, there were ten total posts by the student. And as the final weeks of the course progressed, the student maintained active and “deep” participation in the course.

Figure 3 depicts an evaluation of the effects of teacher solicited engagement on individual student engagement during the baseline (A) observations, the treatment (B) observations, and the newly observed baseline (A) behavior.

![Figure 3. Representation of Increase in Student Posts](image)

The average rating for the baseline phase was .66, the treatment phase was 3.33, and withdrawal phase was 4.66. So, there was an increase in student engagement for the in-
tervention phase followed by an increase in the withdrawal phase. The association between the numbers of postings would support the inference that teacher intervention tended to increase student participation in the discussion board.

However, the overlapping of the number of posts between the treatment and the return to the baseline phase (weeks 4, 5, and 8) would reduce our confidence in such an inference. Further, another limit of this type of inference would also be the possible presence of confounding variables.

Another pleasant result, which was unexpected, was that my presence in calling each student by name, and my interaction with Mary resulted in increased active participation largely by the rest of the class. What was intended to help one student become engaged actually increased the activity of the entire class in the discussion board.

**Discussion**

Although the course syllabus stated the number of posts required for participation and the grade weight associated with those posts, it was apparent that only stating guidelines for participation are not sufficient for some. Palloff and Pratt (2007) suggest that buy-in from the participants is essential in an online course. Participants must first agree to minimum participation standards and understanding what they are committing to. Minimum levels of participation should be “established and agreed upon in order to create a high level of discussion” (p. 21). As seen in this study, one possible way of establishing these standards of participation is to actively and directly engage those students who are at risk of low participation. Also, as seen in this study, when a professor attempts to increase the engagement of one student, it can have an overarching positive impact on the rest of the students. Further, not only is student teacher interaction increased, but student-student, and student-content interaction is also increased. As Stanford-Bowers (2008) states, “an ideal online learning environment is highly interactive with all participants consistently involved with content, the facilitator, and each other” (p. 40). One of the most influential models that deals with interactions in an online course is the one developed by Moore (1989). Moore categorizes interactions into three types:

1. **Learner-Content** - “learner-content interaction is the interaction that occurs between the learner and the content or subject of study” (p. 2). The learner’s interaction with the intellectual content can affect learner’s understanding, the learner’s perspective, or the cognitive structures of the learner’s mind” (p. 2).

2. **Learner-Instructor**, Learners consider learner-instructor interaction highly desirable and necessary (Moore 1989). Interactions involve motivation, feedback, and dialog between the learner and the instructor. After planning a curriculum, the instructor should try to motivate the students and maintain their interest in the subject matter.

3. **Learner-Learner interactions** - Moore states that “learner-learner interaction deals with the exchange of information, ideas, and dialog among learners: be-
tween one learner and other learners, alone or in group settings, with or without the real-time presence of an instructor “(p. 3). The purpose is to share information and ideas for problem solving as a group. Whenever more than two learners interact, group interaction occurs. Successful interaction depends on considering the learner’s age, experience, and level of autonomy”. (pp. 131-132)

4. Hillman, Willis and Gunawardena (1994) added a fourth type of interaction called learner-interface interaction, which occurs when learners use technologies to communicate about the course content, ideas, and information with the instructor and their classmates. Hillman et al, define learner-interface interaction as “the interaction that takes place between the learner and the technology. To be successful, this interaction requires the learner to operate from a paradigm that includes understanding not only the procedures of working with the interface, but also the reasons why these procedures obtain results” (p. 34). As distance education programs use more advanced technologies, the learner’s interaction with those technologies becomes increasingly important. The four types of interactions are described in Table 5 below, and as this study supports, still very relative to online learning in 2011.

Table 5. Four Types of Interaction

<table>
<thead>
<tr>
<th>Interaction Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learner-Content Interaction</td>
<td>The interaction that occurs between the learner and the content or subject of study</td>
</tr>
<tr>
<td>Learner-Instructor Interaction</td>
<td>Interactions involve motivation, feedback, and dialog between the learner and the instructor.</td>
</tr>
<tr>
<td>Learner-Learner Interaction</td>
<td>The purpose is to share information and ideas for problem solving as a group. Whenever more than two learners interact, group interaction occurs.</td>
</tr>
<tr>
<td>Learner-Interface Interaction</td>
<td>The interaction that takes place between the learner and the technology. To be successful, this interaction requires the learner to operate from a paradigm that includes understanding not only the procedures of working with the interface, but also the reasons why these procedures obtain results” (p. 34).</td>
</tr>
</tbody>
</table>

From this study, educators who teach online can take away several important tips or best-practices to engage students in online discussions. Perhaps behaving in these best-practices will avoid the issues of student disengagement in online course. Below is a list of the best practices suggested as a result of this experiment, and following each one is a more detailed discussion based on the literature.
**Best Practices for Discussion Board Engagement**

1. Clearly communicate your expectations regarding student participation in discussions in the syllabus and in an email that is personalized. Detail how often you expect students to read the posts, how often you expect them to post messages, and how you want them to post, such as replying to other students' messages as well as posting their own original messages.

Discussion: Clearly defining expectations for online discussion is important in fully realizing the potential to create an exchange of information and ideas. According to Bender (2003), these must include the expected frequency of participation, how participation counts in the student’s grade, the style of online responses, and the rule of civility.

2. Closely monitor students’ participation, particularly in the beginning of the course, giving encouragement, guidance, and advice to those who are not participating in discussions.

Discussion: It is important to recognize early in the course when students are not participating or meeting the expectations that have been set for discussions, especially if these are a large portion of the students’ grades. Attendance is easy to monitor in a face-to-face course, but is very difficult in an online course and its components. According to Beaudoin (2002), in an online course the potential for cognitive non-attendance is higher. Unless video is used it is not possible to see how attentive students are. In face-to-face interactions an instructor can monitor cognitive attendance by observing a student’s gaze, facial expression, and general demeanor. However, a student could easily log on to an online class and spend his time watching television. Cognitive attendance is more difficult to measure than physical attendance. Thus, most studies focus on physical attendance and make the assumption that it equates with some form of cognitive attendance. It is possible to either mandate physical attendance, or to provide an incentive by including a participation grade based on attendance. It is more difficult to mandate or create incentives for active participation in learning. Beaudoin (2002) states, “Most literature on participation focuses on the reasons that students do or do not participate. In distance courses much or all of the interaction is asymmetric, and participation can be defined by activities such as logging in to symmetric class sessions, participation in discussion boards and forums, and reading class materials.” (p. 316). Thus these are the important ways to monitor students’ attendance in an online discussion.

3. Email students individually who are not participating and show that you are concerned and that you are truly interested in their success and that you are willing to help them all you can, but they must “help you help them”.

Discussion: If after making the expectations clear, and then monitoring the participation of the students, and the level of participation is still not satisfactory for
engaging in the topic, the teacher can try to lure the students in by “a friendly word privately expressed through email, phone, or in person” (Bender 2002, p. 74). Students approach the learning environment with a variety of learning styles and personality traits and thus may respond or feel comfortable with various activities at different rates than other students. Fisher (2001) does however inject the idea that students should be responsible mature adults who should be aware of the consequences of not participating and asks the question, “If students continue to “lurk”, should we continue to try to do anything about this?” So while it is important to initially encourage students’ participation, the question is left to the judgment of the teacher as to how long this should continue so as not to make the student feel targeted.

4. Do not hesitate to offer praise for participation in the discussion board. While some students are more self-directed, some students may find the praise of the professor a strong encouragement to continue in the acceptable behavior. When awarding praise, address why a comment was good, rather than just praising the student. For example, "I like how you incorporated what you learned in social studies into your answer about this story," is better than, "Excellent!". However, either can be very beneficial when giving feedback or praising students in the discussion. Learning is not just cognitive, but also affective, in the sense that it can produce an emotional reaction (Vella, 1997; Fardously, 2001). So, feedback to students likewise can be both. So, giving the student a compliment and then telling them why can have the most benefit to encouraging their participation in the discussion.

**Conclusion**

This was an experiment that was designed to increase the discussion participation of a single student in an online graduate course. As previously stated, just because the number of posts increased, does not increase the quality of the posts. However, the results of the experiment yielded much more benefit than was expected. Not only was learner-instructor interaction increased, but learner-learner, learner-content, and learner-interface interaction were also increased. Additionally, there were broader implications that can serve to inform the practice of all faculty who teach and facilitate online learning. First, enough cannot be said about the importance of noticing early-on the “red-flags” of a student who may be struggling and falling behind in an online discussion. This study showed that with creative thought and energy, a professor can and should be proactive and intervene before a student falls irreversibly behind in a course.

**Further Research**

More of these A-B-A single subject experiments need to be done to affect change in the behavior of students in online courses. Further, the results of several such studies can be compared to recognize difference in student responses to the “treatments” in terms of age, learning styles, prior experience, and level of autonomy. The results can then further the importance of planned, deliberate, teacher interaction in the discussion board and in
online learning as field of study. This could also serve as a springboard for future, larger and more quasi experimental designs that might include a mixed-methods study with qualitative open-ended questions that would help add to the meaning and understanding of the “treatment’s” impact on the learner.

References


©2011 All rights reserved


