A PERSPECTIVE ON STUDENT EVALUATIONS, TEACHING TECHNIQUES, AND CRITICAL THINKING

Prashant Tarun
Missouri Western State University
Craig School of Business
St. Joseph, Missouri

Dale Krueger
Missouri Western State University
Craig School of Business
St. Joseph, Missouri

ABSTRACT

In the United States System of Education the growth of student evaluations from 1973 to 1993 has increased from 29% to 86% which in turn has increased the importance of student evaluations on faculty retention, tenure, and promotion. However, the impact student evaluations have had on student academic development generates complex educational issues. These issues involve teaching critical thinking skills, teaching to the student evaluations, types of tests, grade inflation, student interest in the subject matter, and a student’s sense of entitlement. To avoid the moral and ethical issues associated with educational development and student evaluations, this research compared multiple choice and essay exams as well as comparing an existing student evaluation instrument with another student evaluation instrument. The purpose of this research is to explain the impact of different types of tests with different types of subject matter in an attempt to clarify and reduce distortions, and biases associated with a system of learning that encourages academic development.

INTRODUCTION

This paper covers several important aspects of learning in the United States: type of tests, the critical thinking associated with the tests and the impact of student evaluations on evaluating faculty for promotion and tenure. The introduction first addresses the type of tests and secondly proceeds with the regulations and impact that has developed regarding how to regulate and interpret student evaluations.

First, in the United States multiple choice tests have become heavily used, which raises the question whether multiple choice exams are used too extensively (Phelps, 1996). These exams consist of a stem and a set of options or answers that the person taking the exam can choose the option that has the correct answer called a key and the incorrect answers called distractors (Kehoe, 1995). This type of test does not require the teacher to interpret answers, which helps eliminate teacher bias (DePalma, 1990). The advantages pertain to limited types of knowledge that allows for one answer, which limits testing to lower-order subject matter that has a specific structure. Subject matter that involves problem solving and higher-order reasoning skills are better suited using the essay. Essays are used to judge the comprehension of the material which requires the student to write their answers in an organized presentation.

The essay takes on a number of different forms and styles. The cause and effect requires a causal chain that connects ideas. Categorization breaks ideas into smaller parts. The comparison and contrast analyzes differences between concepts and ideas whereas the descriptive essay provides details usually associated with emotional, physical and intellectual state of the topic. The dialectic and critical essay focuses on an argument or supports a position and usually has examples to clarify a position of strategy. The last two dialectic and critical are usually utilized in Strategy Management classes.

Second, for student evaluations the State of Missouri Legislature passed a law requiring all state colleges and universities to post all student evaluations for all faculty members. Therefore, eliminating student evaluations was not an option at Missouri Western State University or within the Craig School of Business. To assess and improve the use of student evaluations a committee was formed in the Craig School of Business to develop a more in depth perspective on how to interpret student evaluations. At the same time the first step was to develop a new student evaluation instrument that had greater validity and reliability. The second step was to analyze the differ-
A PERSPECTIVE ON STUDENT EVALUATIONS, TEACHING TECHNIQUES, AND CRITICAL THINKING

Prashant Tarun & Dale Krueger

The seven consistently rated the instructor at 31; and the students. Seven students filled in evaluation forms. Two of them for 2.5 grade point average or an ACT of 19. After admission students have to declare a major in one of the four disciplines: accounting, finance, marketing, and management. The initial step in analyzing student evaluations was very straightforward. There are occasionally “outliers” or rogue respondents in college classes who demonstrate no interest in accuracy or fairness in student evaluations of teaching. Anyone who has taught for a while understands that can probably see that student evaluations done in 20 seconds or less that had all “5’s” or all “1’s” [whatever the lowest mark was] for every question. In small classes counted at full value with the others, they tend to bring down average accuracy or fairness in student evaluations of teaching. Respondents in college classes who demonstrate no interest straightforward. There are occasionally “outliers” or rogue respondents who were determined to “punish” instructors for various, frequently [but not always] “imagined” slights or transgressions (Greenberger, 2008).

To present on student evaluations research other variables into some type of context and framework, a review of the literature on educational progress grade inflation, student interest in subject matter, critical thinking and the type of subject matter, perception of students toward left brain and right brain subjects, student assessment about the difficulty of obtaining a grade in various courses, and the implications and suggestions for evaluating student evaluations was undertaken.

This study attempted to compare two evaluations instruments: one based on principles of the Missouri Western State University and a newly designed instrument that incorporated various aspects student learning (critical thinking) along with questions that hopefully provided more appropriate criteria on improving the reliability and validity of student evaluations of the instructor. In addition multiple choice tests and essay tests were compared between Strategic Management and Principles of Management that permitted an analysis associated with an integrated system of learning.

REVIEW OF THE LITERATURE

In 1981, the National Assessment of Educational Progress identified critical skills that workers will need to survive in the 21st century: “Skills in reducing data, interpreting it, packaging it effectively, documenting decisions, explaining complex matters in simple terms and persuading others, to answer how many exams there have been or when did the instructor pass out the syllabus for the course may not answer the question accurately. If the factual questions are not accurately answered this cancels the reliability of the respondent to questions about the pedagogy of the course. Again, this was the initial hypothesis. Another hypothesis was that the use of responses only from respondents who were at least approximately correct on the factual questions would not affect the scores for most instructors. We did not have a firm grasp regarding this second hypothesis. As a result our recollections had been limited. However, outliers who were determined to “punish” instructors for various, frequently [but not always] “imagined” slights or transgressions (Greenberger, 2008).

Students who deal with student evaluation criteria and admissions conflicts for colleges and universities administrators and faculty. An important question that should be addressed is whether educators are focusing their efforts on addressing educational improvement or have rather adapted their tests, courses, and classroom demeanors to improve their student evaluation numbers? The research points toward faculty pandering to modern students’ sense of entitlement. This sense of entitlement appears to be widespread, and depending upon the amount of administrative pressure placed on faculty to generate “good” evaluations, the amount of pandering appears to be influenced by a number of subjectively against the use of student evaluations for retention, tenure, and promotion (Baldrige & Blattner, 2003; Green, Calderon, & Reider, 1998).

Studies that deal with student evaluation criteria and admissions conflicts for colleges and universities administrators and faculty. An important question that should be addressed is whether educators are focusing their efforts on addressing educational improvement or have rather adapted their tests, courses, and classroom demeanors to improve their student evaluation numbers? The research points toward faculty pandering to modern students’ sense of entitlement. This sense of entitlement appears to be widespread, and depending upon the amount of administrative pressure placed on faculty to generate “good” evaluations, the amount of pandering appears to be influenced by a number of subjectively against the use of student evaluations for retention, tenure, and promotion (Baldrige & Blattner, 2003; Green, Calderon, & Reider, 1998).

Prashant Tarun & Dale Krueger
the findings did not show measurable improvement for college students. Over four years of college work 36% of the students did not show improvement in learning, which is perhaps traced to the time spent in academic pursuits. The study indicated students spent less than thirty hours per week on academics, and seniors had not completed a course with 28 or more pages of writing in a previous semester. However, there were differences in majors. Liberal arts students had somewhat higher gains in critical thinking, reasoning and writing compared to students in business, education, social work, and communication. What was significant was the time spent studying alone: five hours. The Arum and Roksa study indicated studying alone was more effective than collaborative learning (Arum and Roksa, 2011).

**METHODOLOGY**

This project included the designing and piloting an alternative student evaluation instrument. The process was to incorporate five factual questions into the instrument. This approach embraced the idea that if these factual questions were correct, then the remaining questions within the student evaluation instrument would improve validity and reliability. For example, using the Missouri Western State University evaluation instrument the student who was not disengaged from the class as to be unable to answer how many exams there have been in the course was to incorporate five factual questions into the instrument.

**Null Hypothesis:**

There is no difference between the instructor’s overall teaching effectiveness rating for a class obtained using the old survey instrument and teaching effectiveness and the likeability rating for the new instrument supported the null hypothesis and produced no statistically significant difference between the two evaluation instruments. There was no significant difference between the old instrument and the new instrument.

Although the Mann-Whitney helped to analyze the Likert scale questionnaires further statistical procedures were tested for association patterns (co-linearity) between the 25 questions on the new instrument. To test for association patterns between survey questions Chi-squared (non-parametric) was used. There were 209 surveys giving a total of 209 student in 11 classes taught by three different instructors. The survey had 25 multiple choice questions. For each of these questions, the answer choices were entered as numbers 1, 2, 3, A, and 5 for choices a, b, c, d, and e, respectively. The instrument was designed to contain 6 embedded “fat” questions that we intended to use on a preliminary sort to eliminate those students whose course involvement was so tenuous as to prevent them from answering what we thought of as simple questions of fact relating to the course.

To test for likeability question (numbered 18) asked the students to respond to this statement: “Indicate your agreement with this statement: ‘I like the instructor for this course.’” When we checked on association patterns using the Null Hypothesis that there was “No association between two variables (or questions),” it was discovered that Q 18 was associated with almost two-thirds (15 of 24) of the questions. Therefore, question 18 determined the overall average for the instructor. To explore this association link between questions further analysis was required.

On this next round of analysis the results were broken down by instructor and the classes they taught in spring semester 2009. To maintain anonymity, instructor names and classes were not identified in this report. Instead, we assigned arbitrary numbers to the instructors and to the classes so that “11” represented instructor #1 class #1; “12” designated instructor #1 and class #2, and so on and so forth. Each student’s answers to the fact-based questions: Q1, Q3, Q4, Q13, Q21, and Q25 were evaluated to identify students that did not agree with the answers picked by majority of students in the class. If there were differences between some student evaluations and the majority of student answers, then class/instructor evaluations become skewed by students who do not display a basic level of class awareness or participation so as to get their facts right about the classes they are taking. In order to support our hypothesis that association patterns exist with student evaluation instrument, various scenarios were constructed for each of the eleven instructor-class combinations based upon students’ answers to the face-based questions Q1, Q3, Q4, Q13, Q21, and Q25. The next step was to explore how answers to the fact-based questions might have been influenced by answers to the likeability question (Q18).

Out of eleven classes, only one instructor in one class [Instructor #1: Class #1] had consistently lower likeability ratings, when students were unable to answer the factual questions correctly. They were excluded from the calculation. For the other ten classes, when non-attentive students were excluded, the evaluation of teaching scores improved. If student evaluations scores and “attentiveness” were independent, the expectation is that 5 or 6 of the 11 classes would have higher student evaluation scores when non-attentive students were included and the other 6 or 5 would have lower student evaluations when non-attentive students were included. Obtaining a 10 to 1 outcome from 11 tries of a 50/50 event is possible, of course, but only 67 times in 10,000 probable. (Chi-Square P=.006656)

In other words, there is both descriptive/intuitive and statistical evidence suggesting a correlation between stu-

---

**TABLE 1: RESULTS FOR OLD SURVEY INSTRUMENT**

<table>
<thead>
<tr>
<th>Instructor</th>
<th>Class Size</th>
<th>Instructor’s Teaching Effectiveness Rating for the Entire Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>28.2.214 0.971</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>1.6 0.699</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>3 1.333</td>
</tr>
<tr>
<td>4</td>
<td>32</td>
<td>2.6875 0.965</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>1.4 0.699</td>
</tr>
<tr>
<td>2</td>
<td>28</td>
<td>2.321 1.89</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>1.923 0.954</td>
</tr>
<tr>
<td>4</td>
<td>19</td>
<td>1.444 0.705</td>
</tr>
<tr>
<td>2</td>
<td>24</td>
<td>1.79 0.93</td>
</tr>
<tr>
<td>3</td>
<td>26</td>
<td>1.3846 0.571</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>1.25 0.452</td>
</tr>
</tbody>
</table>

1- Exceptional, 2- Average, 3- Below Average, 4- Fair, 5- Poor

---

**TABLE 2: RESULTS FOR NEW SURVEY INSTRUMENT**

<table>
<thead>
<tr>
<th>Instructor</th>
<th>Class Size</th>
<th>Instructor’s Likeability Rating for the Entire Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>29 2.24 0.98</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>1 1.2</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>2.64 1.21</td>
</tr>
<tr>
<td>4</td>
<td>32</td>
<td>2.53 1.08</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>1 1.32</td>
</tr>
<tr>
<td>2</td>
<td>29</td>
<td>2.17 1.19</td>
</tr>
<tr>
<td>3</td>
<td>16</td>
<td>2.31 1.19</td>
</tr>
<tr>
<td>4</td>
<td>18</td>
<td>1.33 0.485</td>
</tr>
<tr>
<td>2</td>
<td>17</td>
<td>1.41 1</td>
</tr>
<tr>
<td>3</td>
<td>26</td>
<td>1.19 0.4</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>1.18 0.4</td>
</tr>
</tbody>
</table>

1-Strongly Agree, 2-Agree, 3-Neutral, 4-Disagree, 5-Strongly Disagree
A PERSPECTIVE ON STUDENT EVALUATIONS, TEACHING TECHNIQUES, AND CRITICAL THINKING

By excluding a student's set of responses because the student decides that he/she does not like an instructor, we fail to reject \( H_0 \) since 121 > 59. Also, since \( p = 0.47 \) is greater than 0.05, we fail to reject \( H_0 \).

At Missouri Western State University the original evaluation instrument has indications of co-linearity or association patterns. For example, on question five “The instructor presents the course material clearly and understandably” the evidence indicates that if the students rate the instructor between 2.0 and 2.5 on this question the overall evaluation average will be between one and two. If the students rank the instructor 2.5 to 3.0 the evaluation average falls between 2.0 and 2.5. On the new instrument specific questions number 10 and 11 address critical thinking. Question 10 asks, to indicate your agreement with this statement: I like assignments and exam questions when the answers can be readily checked in the book.” The percentage of students that strongly agreed with the statement was 45.45% and the other five answer percentages were agree at 34.35%, neutral at 17.70%, disagree at 4.44%, and strongly disagree 9.60%. In contrast to question 10 the next question number 11 than asked the students the following: Indicate your agreement with statement: “I like assignments and exam questions whose answers allow for interpretation and creativity”. The percentage of students that strongly agreed with this statement was 10.53% and the other answers were as follows: agree 27.75%, neutral 34.93%, disagree 15.79%, and strongly disagree 11.00%. Question 10 focuses more on courses that are structured with facts and the other question focused more on courses that are structured with facts and different types of courses with different critical thinking objectives and behavioral terms (Gronlund, 1978), which indicates different courses frequently require different levels of critical thinking based on different levels of difficulty. Comparing one instructor with another given the many different types of courses with the different critical thinking levels and different educational objectives becomes an administrative issue. However, if the typical administration/bureaucrat could get past student evaluation averages, student test scores, type of tests that produces differences in critical thinking then business school quality could increase. For example, the Graduate Management Admission Council for Business Schools is now testing for integrative reasoning (Dammon, 2011), and a recent article in Business Education suggests a new rating system for business schools that focuses on quality and learning improvement (Rubin and Morrison, 2015).

Differences in testing procedures and the quality of student products differ precisely in student evaluations between faculty members and also between classes for a single faculty member. These differences aggravate the evaluation problem. Multiple choice exam questions differ from essay exams; and end of chapter essays may reflect specific concepts in the chapter, but may be limited because they usually do not compare and contrast different concepts or ideas. As a result of testing differences and the different types of students enrolled in each class, we find differences in student evaluations not only between classes and between instructors but also between sections of the same class for the same instructor. Although there are differences, this research did not produce statistically different student evaluations between courses and instructors. Current student evaluation procedures are, thus, not reliable for promotion and tenure.

By adding the percentages of the newly student evaluation procedures are, thus, not reliable for promotion and tenure. This educational dilemma between student evaluations and critical thinking is further complicated by the hundreds of different courses offered by the typical university that present a smorgasbord of critical thinking levels for students depending on the nature of course materials and teaching methodologies. Historically, Bloom classified different critical thinking levels in the cognitive domain (Bloom, Engelhart, Furst, Krathwohl, 1956). These cognitive domain classifications start with knowledge and then proceed in the following order: comprehension, analysis, synthesis, application, and evaluation. To expand Bloom’s famous taxonomy of educational objectives, Gronlund divided Bloom’s cognitive domain into instructional objectives and behavioral terms (Gronlund, 1978), which indicates different courses frequently require different levels of critical thinking based on different levels of difficulty.


dents’ ability to answer factual questions in a class and instructor’s likeliness in that class. The statistics based on our scenario analysis supports our hypothesis that the students who are unable to answer factual questions satisfactorily correctly/correctly tend to give lower likeliness ratings to the instructor.

By excluding a student’s set of responses because the student was not able to answer all six fact-based questions correctly the mean composite student evaluation score (independent for the instructor improved and the standard deviation for the class became smaller (indicating more consensus on teaching effectiveness). Apparently, the line between fact and opinion is blurred when an undergraduated student decides that he/she does not like an instructor. The importance of this for our argument is that if non-attentive student responses about whether a syllabus was handed out cannot be relied upon, then their assessment of the instructor’s value in helping to clarify difficult material must be at least suspect.

### Table 3

**SUMMARY OF HYPOTHESIS TESTS:**

<table>
<thead>
<tr>
<th>Instructor</th>
<th>Class</th>
<th>Hypothesis Test Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>486</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>282</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>282</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>282</td>
</tr>
</tbody>
</table>

**Table 3**

**STUDENT EVALUATIONS AND CRITICAL THINKING IMPLICATIONS**

This conflict between student evaluations and student academic development has frequently had a negative impact on both academic skills and the social maturity that college graduates manifest. Self-confidence and self-respect may be seriously jeopardized. If a faculty member attempts to provide instruction that stimulates critical thinking and to construct examinations that actually measure student progress, such a faculty member will probably encounter a significant obstacle when it comes to the student evaluation process. When other variables are added to the mix such as cultural diversity, testing differences (types of tests), grades, brain preference, size of class, critical thinking differences, subject matter differences, different levels of preparation for higher education, anyone attempting to develop a student evaluation instrument that is fair and that provides valid feedback has an enormous challenge with the interaction of the numerous variables that play a role in student evaluations..
A PERSPECTIVE ON STUDENT EVALUATIONS, TEACHING TECHNIQUES, AND CRITICAL THINKING

Prashant Tarun & Dale Krueger

Journal of Learning in Higher Education

Fall 2016 (Volume 12 Issue 2)

ABSTRACT

The research substantiates that student evaluations have inadvertently overtime increased grades in higher education. This study provided evidence on how difficult it is to determine if a particular instrument and how to replace student evaluations into a context. What we have is a conflict with student evaluations grades and the need for faculty in higher education to focus more on developing students. To further this development additional Strategic Management Classes were compared using different teaching techniques and different testing techniques and the results indicated differences in grades and teaching techniques.

For Strategic Management there are eight Essay Questions for first exam. Porter’s buyer and supplier power, competitive rivalry, Deming Quality Management, Barriers to Entry, Economic Leagues, Corporate Strategies, and an Econ Forecast. The second essay exam questions focus on International currency exchange rates including implications, forecasting models, Strategic Alliances and joint ventures etc., BCG Matrix, Different Organizational Structures, Company Cultures, Motivational Practices, and Global and Multinational Strategies. The instructor’s lectured on explaining in detail each of the eight questions, and these are the eight questions that the students are required to take notes and then write out answers for each of the eight questions for ten points. These essays are turned in before each of the eight exams. The instructor teaches to the exam questions and reviews one week before the exam so the students can make adjustments to their answers. By using essay tests that have an extensive writing applications in Strategic Management, the student exam scores improved between the first and second exam with the grade scale at 90% for an A, 80% for a B, 70% for a C, 60% for a D, and were required of the semester with the individual case studies and the group case studies the group exam at the end of the semester averages between 2.5 and 3.0.

However, in the Principles of Management course the multiple choice exams not only decreased with each exam, but also the grades at the end of the course had lower than the first exam scores: 85% for an A, 73% for a B, 63% for a C, and 51% for a D. Even though the multiple choice exam questions were varied, the five exam questions per test for a total of four tests. All the test questions were taken from the test bank, but twenty percent were ranked as easy by the test bank, and the other forty percent were split between moderately and with ten percent considered difficult. Before the exam the instructor reviewed the fifty multiple choice exam questions. As for the teaching methodology for the Principles of Management classes relied simply on the 125 questions per chapter and the test covered three chapters including the final. The final did not have questions over previous chapters. For teaching the textbook power point was utilized...

What follows are the exam results for Strategic Management and Principles of Management. The first column represents the first Strategic Management essay exam average. The second column is the second test average (only two exams) and then the average percentage change between Exam one and Exam Two is the third column. The Principles of Management course reports three exam scores for multiple choice exam questions.

The different tests between the Strategic Management classes and the Principle of Management classes were different. In Principle of Management classes the multiple choice test grades decrease as the course proceeded from historical information on the first exam into multiple choice concepts on subsequent exams including the final, which again, was not comprehensive. In contrast to multiple choice exams the essay approach in Strategic Management shows improvement from the first exam to the second exam with the grade scale at 90% for an A, 80% for a B, 70% for a C, 60% for a D, and were required of the semester with the individual case studies and the group case studies the group exam at the end of the semester averages between 2.5 and 3.0.
tions are reviewed one week before each exam, the review did not produce an increase in test grades.

For the Principles of Management class course grade point average at the end of semester average was between 2.0 and 2.5 on a five point scale compared to 2.5 to 3 point for the Strategic Management classes.

Why the difference in grades and student evaluation between the two courses? The upper level strategic management course that has abstract and complex concepts that have ten or more perspectives and various applications lends itself to teaching the concepts that the students have to explain, apply and then support. For the Principles of Management Course the power point presents an outline of the subject matter with little course depth and very little conceptual comparisons. The average student evaluations for the Principles of Management Course fell between the 2.2 and 2.6 on a five point scale and average about a half point less than the evaluations in the Strategic Management Course, which fall between 1.5 and 2.2.

Conclusion

This research statistically evaluated two different student evaluation instruments. The statistical results show no differences between the use of one student evaluation instrument compared to the other student evaluation instrument, but the type of tests, grades, interest differences in the subject matter (left and right side of brain preference), course difficulty, and student work load are variables that influence the student evaluations averages.

The essay exams in Strategic Management improve from the first exam to the second exam, and the grades for case studies usually avoids any grade below a C whereas the use of test bank multiple choice questions have a detrimental effect on grades in the Principles of Management Course. The grades decrease as the course progresses from exam to exam. By lowering the grade scale in Principles of Management Course, which fall between 1.5 and 2.2.

What this research emphasizes is the type of subject matter determines the type of testing. Courses that are specific and procedural can be taught using multiple choice exams. For example, in a 1994 journal article it was found that in lower level micro and macroeconomics courses, there was not difference between essay exams and multiple choice exams. For example, in a 1994 journal article it was found that in lower level micro and macroeconomics courses, there was not difference between essay exams and multiple choice exams (Walstad and Becker, 1994). More recent research proposes constructed response questions in addition to only multiple choice questions for computer modeling and computer language programing (Simkin and Kuechler, 2005). Further research supports the student preference for multiple choice exams, but also, demonstrates that when students are prepared for the essay exam they appreciated the fairness and validity of the essay exam (Parmenter, 2009).

Courses that lean toward conceptual abstraction require a higher critical thinking approach such as synthesis, where the student is required to compare and contrast the different conceptual alternatives and select the best alternative and support the alternative... The research on comparing the two student evaluation instruments shows no statistical difference between each instrument, but illustrates numerous variables that can affect student evaluations scores such type of test, interest in the subject matter, brain preference, grades, class sizes, etc. However, the research also indicates that matching teach
A PERSPECTIVE ON STUDENT EVALUATIONS, TEACHING TECHNIQUES, AND CRITICAL THINKING

Prashant Tarun & Dale Krueger

REFERENCES


