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 from1973 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-

ences between multiple choice tests and essay tests. The third step was to provide information on a comprehensive system of learning associated with business courses. . Because of the different business disciplines, this study analyzes grades, type of tests, types of students (left or right side of brain, class size, different types of courses (lower and higher level), different critical thinking levels, different course materials, various teaching methodologies and student perceptions.

The research for student evaluations was done at Missouri Western State University, a regional university with 6000 students. It is an open door university broken down into three separate schools: Liberal Arts, Professional Studies, and The Craig School of Business. Within each school various departments set their own admission requirements. For admission into The Craig School Business students have to have a 2.5 grade point average or an ACT of 21. The average ACT in the School of Business is 21- 22. 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 straight forward. 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 decade or so can probably recall student evaluations done in 20 seconds or less that had all "5's" or all "e's" [whatever the lowest mark was] for every question. In small classes when these are counted at full value with the others, they tend to bring down average scores significantly. For example, in one 400 level evening class that Professor M had at this school, there were 8 students. Seven students filled in evaluation forms. Two of the seven consistently rated the instructor at 3[1]; and the other five were mostly 1 and 2 ratings. The result was that the instructor had an average rating of 1.82 with a 0.78 standard deviation. Without the two outliers, the instructor would have an average rating of approximately 1.40. The 1.4 would place the instructor in the top half of instructors university-wide (mean = 1.555) and still higher in the school of business (mean = 1.894). The question is whether the differences are statistically significant to warrant a decision on who is the better teacher? According the statistical research the statistical significant research can be strong or weak and small or large. For example, the difference between student evaluations of 1.90 and 1.94 at a significant level of .05 with a standard deviation of .8 requires a sample size of 3074 using a Z test of independent samples (McClave & Benson, 2008).

Despite the statistical difficulty of measuring student evaluations we proposed to pilot an evaluation instrument that would contain five or so factual questions. Formally, we hypothesized that the "rogue respondents"[2] or "outliers" would answer the factual questions accurately. Privately the speculation is that this might not be true. If this hypothesis were false and the speculation true, then an initial sort to screen out student responses on the conduct of the class that were factually wrong should make the remaining evaluations more reliable. For example, a student who is so disengaged from the class as to be unable 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 to outliers 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 the present one in use at 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 test results 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" (NAEP, 1981). These skills point toward the need for colleges and universities to identify and develop students' abilities to "to turn facts into concepts, to turn concepts into a policy or plan, and to see the issue and define the problem within a problematic situation" (Flower, 1990). Since 1981 periodically attention has been drawn to adult literacy and the problem associated with workers that do not have the ability to perform work tasks that are increasing becoming more complex and technical. The problem isn't people who can't read and write, but those who read and write at lower levels than the task demands (Grimsley, 1995). Despite the attention to the goal of improving our educational system and concomitantly the skills of our students not much progress has been made by our educational system other than articles on how to make our schools better for our children (Symonds, 2001).

The goal of developing critical thinking skills in students and the goal of improving student evaluation numbers in higher educational institutions has generated moral and professional conflicts for college and university 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 substantiated by a number of studies against the use of student evaluations for retention, tenure, and promotion (Baldwin and Blattner, 2003; Green, Calderon, and Reider, 1998).

Studies that deal with student evaluation criteria and administrative cognitive processes in performance appraisal that were conducted in field settings raises questions about the usefulness of this practice. Despite the lack of reliable and valid information business schools use the evaluations for a number of purposes (Cleveland, Murphy, Williams, 1989). The results of these evaluations are used for various human resource decisions. However, if the objectives in the evaluation instrument are unclear and the criteria measuring those objectives are vague, there will be an unsatisfactory payoff for the employee, the organization, and the evaluative participant. The result can be confusion and misapplication. For example, student evaluations may depend on the context of other students, on previous student performance, the level of student development, the type of subject matter, student's interest in the subject matter, testing difficulty, instructor's knowledge, teacher-student relationships, the teacher's organizational skill, communication skill, and the content difficulty.

To unravel the evaluation process researchers have attempted to design standardized instruments to improve the reliability and validity of the ratings. Unfortunately, there is no substantial evidence to support the fact that student evaluations improve instructional quality (Adams, 1997), and yet the research indicates college instructor's should be measured against seven dimensions: (1) instructor knowledge, (2) testing procedures, (3) student-teacher relations, (4) organizational skills, (5) communication skills, (6) subject relevance, (7) utility of assignments (Robbins, 2000). Although these dimensions have been identified, the problem is universities and colleges have tried to implement classroom evaluations to gather information on students perceptions of what transpired in the classroom during the duration of the course to obtain information for promotion, retention, and feedback. These evaluation instruments have fall short. For example, one aspect of the research indicates non-verbal behavior warmth and supportiveness (interpersonal behavior) are related to the teacher's student evaluation (Ambady & Rosenthal, 1993). However, these dimensions need criteria to support the seven dimensions. For this study two evaluation instruments were compared to provide a possible benchmark and greater understanding of student evaluations and the impact on critical thinking, which includes differences between test multiple choice tests and essay tests.

Because of the emphasis in higher education on student evaluations, grade inflation seems to correlate with the increased use of anonymous semester-ending student evaluations. In 1987 27% of the high school students taking the SAT test had GPA's in the A-plus to A to A minus range, and by 2007 the percentage of "A-students" taking the SAT had increased to 43% (Caperton, 2009). This grade inflation contributes to what students perceive as self-entitlement. This self-entitlement translates into students pressuring professors for higher grades based on their special needs and preferences (Greenberger, Lessard, Chen, & Farrugia, 2008). At the university level recent research has pointed out that student evaluations are positively correlated with grades (Weinberg, Hashimoto, & Fleisher, 2009); many faculty contend that student evaluations play a very significant role in tenure and promotion. Therefore, it is not unusual for faculty to resort to open book exams, more true-false questions, and essay exams that emphasize lower levels of critical thinking to generate higher grades and better student evaluations rather than focus on educational development. Harvard reported "one-fourth of all grades given to undergraduates are now A's and another fourth are A-'s (Mansfield, 2001).

The most recent article on complex reasoning and writing skills (General collegiate skills appeared in the Chronicle of Higher Education on January 20, 2011. (Vedder, 2011). Using the Critical Leaning Assessment (CLA) to measure the gains in critical thinking, reasoning, and writing skills

the findings did not show measureable 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 20 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 is so disengaged from the class as to be unable to answer how many exams there have been in the course would not be able to respond appropriately to questions about the pedagogy of the course. The alternative student evaluation instrument requires students to answer factual questions. If these factual questions were not correct, then the instructor's overall student evaluation ranking would not be correct.

To statistically compare the two student evaluations instruments the null and alternative hypotheses follow:

Null Hypothesis:

***H*$_0$:** There is no difference between the instructor's overall teaching effectiveness rating for a class obtained using the old survey instrument and the instructor's likeability rating for a class obtained using the new survey instrument.

Ho: There is no difference between multiple choice and essay exams

Alternate Hypothesis:

***HA*:** There is a difference between the instructor's overall teaching effectiveness and likeability rating for a class obtained using the old survey instrument and for a class obtained using the new survey instrument.

Three instructors used the new evaluation instrument in the following classes:

- Instructor one: class 1 Management of Organizations, classes 2-4 Strategic Management
- Instructor two: class 1 Advanced Income Tax, classes 2-3 Business Law
- Instructor three: class 1: International Finance, class 2: Finance Principles, classes 3-4: Introduction to Statistics

Ha: There is a difference between multiple choice exams and essay exams

A Comparison of the Two Survey Instruments

Assumptions: In order to compare the old (current university form) with the newly piloted form, we had to make certain assumptions. They were:

1. Likert scales for old and new survey instruments are comparable; and
2. Instructor's likeability ratings from the new survey instrument can be compared with the ratings of instructor's overall teaching effectiveness from the old survey instrument.

To statistically compare the two evaluation instruments the Mann Whitney U Test was utilized.

In evaluating the statistical results using the Mann-Whitney for five senior level courses and six sophomore-junior level courses, the teaching effectiveness for the old 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 given to a total of 209 students 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, 4, and 5 for choices a, b, c, d, and e, respectively. The instrument was designed to contain 6 embedded "fact" 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.'"[3] 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 fact-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 the 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

Instructor	Class	Class Size	Instructor's Teaching Effectiveness Rating for the Entire Class	
			Mean	Std. Dev.
1	**1**	28	2.214	0.917
	2	10	1.6	0.699
	3	10	3	1.333
	4	32	2.6875	0.965
2	**1**	10	1.4	0.699
	2	28	2.321	1.09
	3	13	1.923	0.954
3	**1**	18	1.444	0.705
	2	24	1.79	0.93
	3	26	1.3846	0.571
	4	12	1.25	0.452

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

Table 2
Results for New Survey Instrument

Instructor	Class	Class Size	Instructor's Likeability Rating for the Entire Class	
			Mean	Std. Dev.
1	1	29	2.24	0.98
	2	10	1.5	1.2
	3	11	2.64	1.21
	4	32	2.53	1.08
2	1	10	1.1	0.32
	2	29	2.17	1.19
	3	16	2.31	1.19
3	1	18	1.33	0.485
	2	17	1.41	1
	3	26	1.19	0.4
	4	11	1.18	0.4

1-Strongly Agree, 2-Agree, 3- Neutral, 4-Disagree, 5-Strongly Disagree

Table 3
Summary of Hypothesis Tests:

Instructor	Class	Hypothesis Test Results
1	1	$U = 406$ Critical Value of the Mann-Whitney U test at $\alpha=0.05$ for $n_1=28$ and $n_2=29$, $U_{critical}=282$ We fail to reject H_0 since 406 > 282. Also, since p = 1.00 is greater than 0.05, we fail to reject H_0.
	2	$U = 37$ Critical Value of the Mann-Whitney U test at $\alpha=0.05$ for $n_1=10$ and $n_2=10$, $U_{critical}=23$ We fail to reject H_0 since 37 > 23. Also, since p=0.35 is greater than 0.05, we fail to reject H_0.
	3	$U = 47$ Critical Value of the Mann-Whitney U test at $\alpha=0.05$ for $n_1=10$ and $n_2=11$, $U_{critical}=26$ We fail to reject H_0 since 47 > 26. Also, since p=0.60 is greater than 0.05, we fail to reject H_0.
	4	$U = 572.5$ We fail to reject H_0 since 572.5 > $U_{critical}$ at $\alpha=0.05$ for $n_1=32$ and $n_2=32$. Also, since p=0.42 is greater than 0.05, we fail to reject H_0.
2	1	$U = 39.5$ Critical Value of the Mann-Whitney U test at $\alpha=0.05$ for $n_1=10$ and $n_2=10$, $U_{critical}=23$ We fail to reject H_0 since 39.5 > 23. Also, since p=0.44 is greater than 0.05, we fail to reject H_0.
	2	$U = 362$ Critical Value of the Mann-Whitney U test at $\alpha=0.05$ for $n_1=28$ and $n_2=29$, $U_{critical}=282$ We fail to reject H_0 since 362 > 282. Also, since p=0.49 is greater than 0.05, we fail to reject H_0.
	3	$U = 121$ Critical Value of the Mann-Whitney U test at $\alpha=0.05$ for $n_1=13$ and $n_2=16$, $U_{critical}=59$ 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.
3	1	$U = 156$ Critical Value of the Mann-Whitney U test at $\alpha=0.05$ for $n_1=18$ and $n_2=18$, $U_{critical}=99$ We fail to reject H_0 since 156 > 99. Also, since p=0.86 is greater than 0.05, we fail to reject H_0.
	2	$U = 135$ Critical Value of the Mann-Whitney U test at $\alpha=0.05$ for $n_1=24$ and $n_2=17$, $U_{critical}=129$ We fail to reject H_0 since 135 > 129. Also, since p=0.07 is greater than 0.05, we fail to reject H_0.
	3	$U = 283.5$ Critical Value of the Mann-Whitney U test at $\alpha=0.05$ for $n_1=26$ and $n_2=26$, $U_{critical}=230$ We fail to reject H_0 since 283.5 > 230. Also, since p = 0.32 is greater than 0.05, we fail to reject H_0.
	4	$U = 61.5$ Critical Value of the Mann-Whitney U test at $\alpha=0.05$ for $n_1=12$ and $n_2=11$, $U_{critical}=33$ We fail to reject H_0 since 61.5 > 33. Also, since p=0.79 is greater than 0.05, we fail to reject H_0.

dents' ability to answer factual questions in a class and instructor's likeability 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 tend to give lower likeability 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 (average) 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 undergraduate 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.

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 and 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..

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 1.44%, and strongly disagree .96%. 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 specific procedures such as finance and accounting. Question 11 more on courses that require synthesis for application. Similarly questions 14, 15, 16 on whether the concepts were more interesting, valuable, and difficult did not produce any substantial deviations. However, in reviewing some of the results by subject area, type of tests, and grades there were some differences that indicate student evaluations vary depending on the type of course.

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 with the difficulty increasing 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. 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 administrator/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 students produce differences in student evaluations between faculty members and also between classes for a single faculty member. These differences aggravate the evaluation problem. Multiple choice exams 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 instrument for whether students strong agree and agree on each of the questions associated with student

evaluations by subject, type of tests, grades, size of class, major, produces additional insight on the difficulty and the complexity of interpreting student evaluations fairly. For example, question number 10, "I like assignments and exam questions when the answers can be readily checked in the textbook." The upper level courses MGT 419 Strategic Management, Tax, and International Finance the percentage of students favoring the question 10 was 69 % versus 82% for the lower level classes: Principles of Management, Business Law, Principles of Finance, and Business Statistics.

Question 11 asks students whether they like assignments and exam questions that allow for interpretation and creativity. The average percentage on question 11 for all subject areas was 40% whereas for question 10 where the assignments and exams are tied back to the textbook the average student percentage was 77% for all subject areas.

What is interesting is the difference in percentages for the two classes of MGT 419 one class average for question 11(assignments and exam questions allow for interpretation and creativity) was 60% and the other class was 36%. In checking the number of majors by subject the class makeup was quite different. The class that rated question 11 at 60% has 12 students with nine marketing and management majors and the class that rated question 11 at 36% had 14 students with 10 of the students majoring in finance and accounting. This percentage difference indicates a brain preferences(left or right) may play a role in student evaluations.

Turning to questions 14 and 15 substantial differences exist between the strategic management classes. Question 14 asked whether "the concepts in this course were more interesting than the concepts in most other courses I have taken," and question 15 asked "The concepts in this course were more valuable than concepts in most other courses I have taken". For question 14 the student ranking was 60% for the marketing and management majors and 18% for the finance and accounting majors. On question 15 the percentage difference was 70% for marketing and management majors compared to 36% for finance and accounting. However, on question 23 that asked, "The instructor stimulated my interest in this subject," the class with the marketing and management students ranked question 23 at 90% and the class with the finance and accounting students ranked question 23 at 18%. In short, questions 14 (interesting concepts), question15 (concepts were more valuable), and question 23 (instructor stimulated my interest) the differences were considerable, and yet on question 20 which asked "it was harder to get a good grade in this course than in other courses," there was no difference between the two Strategic Management (MGT 419) classes: 82% compared to 81%. Even though one class had more finance and accounting students and the other class had more marketing and management majors. Then question arises whether the teaching and assignments were different between the two classes? The answer is no. In teaching Strategic Management 419 there was no difference in the lectures, exams, individual case studies, and the group case studies, and all exams and individual case studies were graded anonymously by having the students use an identifying mark that they selected. When the papers were handed back the students wrote their names on the papers, and instructor recorded the grades.

For the other upper level courses International Finance and Tax the concepts were more interesting than other courses (question 14) the percentages were respectively 72% and 50%, but for the lower level courses Principles of Management MGT 305; Business Law, GBA 211; Principles of Finance, FIN 301; and Business Statistics, GBA 210 the average was 33%. On question 15 (concepts in this course were more valuable than concepts in other courses) there was a variance. Principles of Management and Business Law more right brain subjects averaged 39% whereas Tax, Principle of Finance, Business Statistics the more(quantitative and procedural subjects averaged 68%.

Question 16 asks whether "The concepts in this course were more difficult than concepts in most other courses I have taken". The total average for question 16 was 59%. In comparison the tax course ranking was 80%. Question 20 asks whether "It was harder to get a good grade in this course than in other courses". The total average for question 20 for the tax course was 60%. However, the strategy courses were ranked higher at 80% and 82%, which is consistent with question 12 which indicated the strategy course required more work than other courses. Question 23 asks the students does "The instructor stimulate my interest in the subject". The average was 55% with a range of 24% to 90%. For question 23 on whether the instructor stimulated my interest in the course the upper level courses Strategic Management (two classes), Tax, and International Finance scores were 90%, 18% for Strategic Management. The 90% class had a predominance of marketing and management majors, and the 18% class had accounting and finance majors. For the other upper level courses Tax, and International Finance the scores were respectively 80% and 83%. Why the difference in the Strategy classes? To explain the difference between the two strategy classes remember one class was populated with 75% marketing and management majors and the other 75% finance and accounting, and the research indicates most marketing and management majors are right brain whereas finance and accounting majors are usually left brain (Krueger, 2009). . Therefore, brain preference stimulates interest in the subject matter and plays an important role not only in how students evaluate the course and the instructor, but also indicates a strong connection between high student interest in the subject matter, and student learning outcomes (Bergin 1999: Frymier, Shulman, & Houser, 1996: HIDI, 1990; Schiefele, 1991, 1996). According to Schiefele a student's subject matter interest increases learning because subject matter interest encourages student intrinsic motivation. Specific types of tests that represent specific learning strategies that correlate with student interest and motivation lead to student internalization and ownership of material (Dewey, 1913). These connections in turn lead to different levels of critical thinking and can produce differences student evaluation differences, but again not significant statistical differences.

Question 23 on whether the instructor stimulated my interest in the course the lower level courses Principles of Management, Business Law, Principles of Finance and Business Statistics averaged 49%. Why? The lower level classes students usually have not committed themselves to a specific major. Therefore, interest in the subject matter at this level becomes difficult to assess.

IMPLICATIONS

The research substantiates that student evaluations have inadvertently overtime increased grades in higher education. This study provided evidence on how difficult it is to design a better student evaluation instrument and how to place 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 alternative hypothesis 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, Business Strategies, 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 centered on explaining in depth 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. Then the instructor reviews the test questions before the students take the exam. This approach enables the student to prepare for the eight questions and out of the eight three are selected for the test. For the Principles of Management course

the fifty exam questions per test for a total of four tests. All the test questions were taken from the test bank, and twenty percent were ranked as easy by the test bank, and the other forty questions were split between moderate 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 exams.

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 more abstract 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 to the second exam. In the Strategic Management classes the exam questions are handed out at the start of the semester, and the students are given points for developing their answers to the questions before they take the exam. 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 and application approach in Strategic Management, the group 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 60% for a F. At the end of the semester with the individual case studies and the group case studies the group course grade point 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 the teacher at the end of the course had to lower the grade scale: 85% for an A, 73% for a B, 63% for a D, and 51% for an F. Even though the multiple choice exam ques-

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 theStrategic 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 the assumption is the instructor more than likely avoids extreme negative student evaluations. The other variables, subject matter interest, course difficulty may play a role in how students perceive the course, but the important implication is the teaching and learning methodology associated with the subject matter. Whether the subject matter fits the type of test, and requires the student to develop their organization skills, writing skills, and upper level critical thinking skills such as synthesis becomes the important question.

In the strategic management classes what is apparent the teaching techniques illustrate a system of learning that promotes academic Improvement and written about a few years ago (Stefani, 2011).

Table 4
Exam Results
MGT 419 Strategic Management

Exam 1	Exam 2	Percent Change
Fall 14		
69.43	79.15	9.72
6-A	9-A	
6-B	8-B	
7-C	8-C	
3-D	5-D	
10-F	2-F	
Summer 2014		
70.62	74.68	4.06
3-A	4-A	
2-B	2-B	
6-C	10-C	
6-D	D-1	
F-4	F-2 (attendance problem)	
Spring 2014		
75.42	83.56	8.14
3-A	7-A	
12-B	10-B	
5-C	3-C	
1-D	1-D	
2-F	1-F	
Spring 2014		
68.84	74.89	6.04
3-A	4-A	
5-B	5-B	
3-C	7-C	
4-D	1-D	
3-F	2-F	
Fall 2013		
70.33	79.18	8.85
0-A	6-A	
7-B	11-B	
10-C	4-C	
4-D	2-D	
6-F	4-F	

Table 4
Exam Results
MGT 419 Strategic Management

Exam 1	Exam 2	Percent Change
Summer 2013		
72.50	84.30	11.8
4-A	4-A	
3-B	5-B	
2-D	0-D	
4-F	0-F	
Spring 2013 Day Class		
77.42	75.54	-1.88
1-A	1-A	
11-B	10-B	
7-C	4-C	
5-D	D-8	
0-F	3-F	
Spring 2013 Evening Class		
70.17	81.38	11.21
2-A	7-A	
8-B	10-B	
9-C	7-C	
2-D	4-D	
9-F	1-F	

Table 5
Exam Results
Principles of Management

Exam 1	Exam 3	Final
Spring 14		
72.03	72.08	73.67
A-0	A-3	
B-2	B-4	
C-12	C-8	
D-8	D-7	
F-3	F-3	
Spring 2012		
79.2	66.60	71.74
A-3	A-0	
B-8	B-6	
C-9	C-10	
D-7	D-10	
F-0	F-5	
Spring 2012		
79.05	71.45	71.88
A-4	A-2	
B-13	B-7	
C-8	C-8	
D-8	D-10	
F-0	F-5	
Spring 2011		
74.6	71.0	77.88
A-4	A-1	
B-13	B-7	
C-8	C-5	
D-10	D-8	
F-3	F-4	

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 (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 size, etc. However, the research also indicates that matching teach-

ing and learning methodology is far more important than the emphasis that has been placed on student evaluations. A recent article on faculty development suggested different assessment procedures for faculty that focuses on academic improvement (Fink, 2013).

REFERENCES

Adams, J. V., (1997). *Student evaluations: The rating game,* 1 Inquiry(No. 2, fall 1997) 10-16.

Ambady, N. & Rosenthal, R. (1993). *Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness,* Journal of Personality and Social Psychology 64 (no.3 1993) 431-441at 432.

Arum R.and Roksa. J. (2011). *Academically adrift.* http:/www.press.uchicago.edu/ucp/books/book/Chicago/A/bo10327226.html.

Baldwin, T. and Blattner, M. (Winter, 2003). Guarding against potential bias in student evaluations, 51(1) 28.

Bergin, D. (1999) Influences on classroom interest. *Educational Psychologist,* 34(2), 87-98.

Bloom, B. S., Engelhart, M. D., Furst, E. J., and Krathwohl, D. (1956). *Taxonomy of Educational Objectives Handbook* 1: *Cognitive Domain*. NY: David McKay.

Caperton, G. (September, 2009). Test data allow better decisions, *U.S. New and World Report*, 24.

Cleveland, J. N., Murphy, K. R., & Williams, r. E. (1989) Multiple uses of performance appraisal: Prevalence and correlates. *Journal of applied Psychology, 130-35.* B., Calderon, T., Reider, B., (February, 1998). A content analysis of teaching evaluation instruments used in accounting departments, *Issues in Accounting Education*, 13(1) 15.

Dammon, R. (2011). Critical-thinking students at b-school, *Wall Street Journal, A14.*

DePalma, Anthony (1 November 1990). Revisions Adopted in College entrance Tests. (http://www.nytimes.com/1990/11/01/us/revisions-adopted-in-coleege-entrance-tests.html) *New York Times*. Retrieved 22 August 2012.

Fink, L. (2003). *Creating significant learning experience: An integrated approach to designing college courses.* San Francisco: John Wiley.

Fink, L. (2013). *Innovtive ways of assessing faculty development,* New Deriectionsfor Teacing andLearning,no 133 Spring 2013. Wiley Periodicals, Inc.

Frymier, A., Shulman, G. & Houser, M. (1996). The development of a learner impowerment measure. *Communication Education, 45 181-19.*

Green, B. K., Calderon, T. G. and Reider, B. P. (1998). A content analysis of teaching evaluation instruments used in accounting departments, *Issues in Accounting Education*, Vol. 13, issue 1, 15-30

Greenberger, E., Lessard, J., Chen, C., and Farruggia, S. (2008). Self-entitled college students: contributions of personality, parenting, and motivational factors, *Journal of Youth Adolescence,* 37: 1193-1204.

Grimsley, K. D. (1995). Companies struggle with illiteracy in workplace, *St. Joseph News Press,* October 1, p. C1.

Gronlund, N. E. (1978) *Stating objectives for classroom instruction (Second edition*.). NY: Macmillan Publishing.

Herrmann, N. (1992). *The creative brain.* Aracata Graphics, Kingsport, Tennessee.

Hidi, S. (1990). Internst and its contribution as a mental resource for learing. *Review of Educational Research,* 60, 549-571.

Kehoe, Jerard (1995). Writing multiple-choice test itmes (http://PAREonline.net/getvn.asp?v=4&n=9) *Practical Assessment, research & Evaluation,* 4/9. Retrieved February 12, 2008.

Krueger, D. (November, 2009) A cognitive inventory of business students. Kentucky Journal of Excellence in College Teaching and Learning, Vol. 7: 21-28

Langbein, Laura (2008). Management by results: student evaluation of faculty teaching and the mis-measurement of performance, *Economics of Education Review.* August, vol. 27 Issue 4, 417-428.

National Assessment of Educational Progress (1981). *Reading, Thinking and Writing*: *Results from the 1979-81 National Assessment of Reading and Literature*. Denver: Education Commission of the States.

McClave, J. T. & Benson, P. G. (2008). *Statistics for business and Economics.* 10th Edition, 265, 445.

Mansfield, H. C., (2001). *Grade inflation: It's time to face the facts.* The Chronicle of Higher Education, April 6: http://chronicle .com/article/Grade-inflation-its-time-to/9332.

Morgenegg, B. L. (2000, April). Perceptions of "Faculty Teaching Characteristics", Paper presented to the *Colorado Regional Higher Education Assessment Conference*. Denver, CO.

Parmeter, D. A., (2009). *Essay versus multiple-choice student preferences and the underlying rationale with implications for test construction.* Academy of Educational Leadership Journal, Vol. 13, Number 2, 57-72.

Phelps, Richared, (1996), Are US students the most heavly tested on earth? */dedycatuibak Neasyrnebt: Issues and Practice, 15 (3): 19-27m diu:10.1111/j.1745-3992.tb00819.x(https://dxdoi.org/101111%2Fi.1745-3992.1996.tb00819.x)*

Rogers, E. M. (1997). *A History of Communication Study.* New York, NY: The Free Press, 396-399.

Rubin, R. S. and Morgeson, F. P., (2015). *Redefining quality,* Business Education, January-February, 48-51.

Schiefele, U. (1996). Interest, learning, and motivation. *Educational Psychologist,* 26, 299-323.

Schiefele, U. (1996). Topic interest, text representation, and quality of experience. *Contrmporary Educational Psychologist,* 21, 3-18.

Simkin, M. G.,and Keuchler, W. I., (2005). *Multiple-choice tests and student understanding: what is the connection,* Decision Sciences Journal of Innovative Education, Spring , Vol. 3 issue,73-97.

Stefani, L., (2011). *Evaluating the effectivenessof academic development: principles and practice*, Abingdon, England Routledge 222.

Symonds, W. C. (2001, September). A Is for Answers. *Reader's Digest*. 99-105.

Vedder, R. (2011). Academically adrift: a must read, *The Chronicle of Higher Education*, http://chronicle.com/blogs/innovations/academically-adrift-a-must-read/28423?sid=pm&ut.

Walstad, W. B. and Becker, W. E., (1994). *American Economic Review, Vol. 84. Issue 2, 193-194.*

Weinberg, B., Hashimoto, M., and Fleisher, B., (Summer, 2009). Evaluating teaching in higher education, *Journal of Economic Education,* 227-261.