3-26-2015

The Tail Wagging the Dog: An Overdue Examination of Student Teaching Evaluations

Patti Miles

Deanna House
University of Nebraska at Omaha

The Tail Wagging the Dog; An Overdue Examination of Student Teaching Evaluations

Patti Miles[1] & Deanna House[2]

[1] Associate Professor of Management, University of Maine, USA

[2] Assistant Professor of Management Information Systems, Ohio University, USA

Correspondence: Patti Miles, Associate Professor of Management, University of Maine, USA. Tel: 1-207-951-1994 E-mail: patti.miles@maine.edu

Received: February 2, 2015 Accepted: March 23, 2015 Online Published: March 26, 2015

doi:10.5430/ijhe.v4n2p116 URL: http://dx.doi.org/10.5430/ijhe.v4n2p116

Abstract

Purpose: The purpose of this research is to examine the impact of several factors beyond the professor's control and their unique impact on Student Teaching Evaluations (STEs). The present research pulls together a substantial amount of data to statistically analyze several academic historical legends about just how vulnerable STEs are to the effects of: class size, course type, professor gender, and course grades.

Design/methodology/approach: This research is utilizes over 30,000 individual student evaluations of 255 professors, spanning six semesters, during a three year time period to test six hypotheses. The final sample represents 1057 classes ranging in size between 10 and 190 students. Each hypothesis is statistically analyzed, with either analysis of variance or a Regression model.

Findings: This study finds support for 5 out of 6 hypotheses. Specifically, these data suggest STEs are likely to be closest to "5" (using a 1-5 scale with 5 being highest) in small elective classes, and lowest in large required classes taught by females. As well we find support for the notion that higher expected course grades may lead to higher STEs.

Practical implications: The practical significance of this research is important. First this research utilized a large data set spanning several years and hundreds of professors and thousands of students and rigorous statistical analysis to assert several important findings. Indeed STEs are impacted significantly by class type, class size, the gender of the professor and the expected course grade. With these findings we suggest a more comprehensive mechanism is in order for evaluation of teaching effectiveness.

Social implications: This research could have great social implications if widely read across academic circles. Indeed the tail is wagging the dog; or the student is influencing teaching across America's universities. It is time to examine teaching effectiveness through a different lens, because using teaching evaluations to determine promotion and tenure, sparse bonus allocation, and teaching awards may be short sighted.

Research limitations: While this research is statistically accurate, it is limited by the notion that the data was collected from one large area. As such, care should be taken in generalizing these results to other areas that may have different demographic composition, funding etc.

Originality/value: To the best of the authors' knowledge this research is the first of its kind to statistically analyze such a large body of data and provide a useful guide to help evaluate professors utilizing what information is available.

Keywords: Student teaching evaluations, Teaching effectiveness, Gender, Class size, Course grades

1. Introduction

Few topics are debated among faculty quite as fervently as the end of semester Student Teaching Evaluations (STEs), particularly among those whose scores are on the lower range. Most commonly heard are assertions that specific types of classes or situations do not yield as favorable a level of evaluations as do others, and that factors most affecting such evaluations carry little worth with respect to teaching effectiveness. While some of this debate is no doubt stirred by those who would like to place external blame for results that are not to their liking, the preponderance of research over the last 30 years would indicate that such assertions may be grounded in the notion

that instructors are evaluated not on effectiveness, but through a variety of factors well beyond their control. Thus, this research calls into question the validity, reliability, and/or fairness of the use of STE results.

In recent years the use of the STE to evaluate student teaching has become more and more prevalent and research suggests that such evaluations are less and less likely to improve teaching effectiveness, and instead be used to evaluate teaching quality (Knol, Veld, & Mellenbergh 2013). Despite this, the use of "raw" STEs for evaluating and rating teaching appears to be, if anything, rising in recent years and seems likely to remain an important component of the evaluation of faculty. Thus, the present research seeks to provide a modest inquiry into factors that statistically influence STE scores, yet do not reflect teaching effectiveness; and provide some practical advice for appropriate use of STEs when evaluating professors.

The research is organized as follows. The literature review is centered on four factors that may affect STEs without being true indicators of teaching effectiveness; followed by the development of six hypotheses. Next the methods section explains the origin of the sample, the methodological choices, and results. Discussion and interpretation of the analysis culminates in the presentation of several suggestions that could enhance the interpretation of STEs.

2. Literature Review and Hypotheses Development

Student Teaching Evaluations (STEs) began just after World War II (Eiszler, 2002), and have grown in sophistication and prevalence since that time (McKeachie, 1990). This large body of literature has continued to swell, examining both the reliability and validity of the measures (Cohen, 1981; Costin, Greenough, & Menges, 1971), student perceptions of the process, the influence of moods (Marsh, 1984), and the role STEs make in tenure decisions (Backer, 2012; Gray, & Bergmann, 2003, Bray, 1980;). Of note is the research by Eiszler (2002) who conducted a survey of social science deans and highlighted both the pivotal role of STEs in personnel decisions and their role as significant contributors to grade inflation. STEs also play a part in annual pay raises. For example, one study examined 10 years of data across five universities and discovered 45% of faculty pay is determined by teacher effectiveness as quantified on student teaching evaluations (Filetti, Wright, & King, 2010; Stratton, Myers & King 1994). Yet, even in light of this extensive research, STEs remain the most popular tool for making promotion, tenure and pay decisions (Southwell, Gannaway, Orrell, Chalmers, & Abraham, 2010).

As Houston, Meyer, & Paewai (2006) suggest little research provides guidance for administrators on how to effectively utilize STEs to make personnel decisions, while balancing complex teaching loads, classroom size, and shrinking budgets with respect to temperamental faculty desires. Using this background, the present research began with a broad review of the literature and rapidly reveals four factors that historically seem to influence STEs, yet hold little value in assessing teaching effectiveness. The four factors that continually appear in the literature as impacting STEs are class type (core or elective) and class size, professor gender and course grades.

1.1 Class Type (core or elective)

Required classes, such as those in a university or college "core", generally consist of large groups of students with broad interests and grades that are typically lower than in higher level elective classes. As such, the teaching of these courses has been debated among faculty over the years as being 'more difficult' to teach. To that end, several large studies consistently find teachers of elective courses receive more favorable evaluations than their colleagues teaching required classes (Drago & Wagner, 2004; Whitten & Umble; 1980; Pohlmann, 1975; Lovell & Haner, 1955). Researchers have suggested a number of possible reasons for these findings, including student attitudes towards the topic, the quantitative nature of some of the required courses, student perception of quality, the professor's teaching style (Bosshardt, 2001; Marsh & Roche, 1997) and the students interest and engagement (Darby, 2006a); all of which affect STEs. Related research finds STEs to be influenced by a student's year in school; asserting a positive correlation between the two (Ory, 2001). Given that students taking required classes are likely to be at a lower level in school and have less interest in the subject (than those taking a major elective) we assert the following:

H_1: Student teaching evaluations in elective classes will be higher than in required classes.

1.2 Class Size

As economic times remain uncertain and funding for institutions of higher learning continues to decline, many universities are allowing class size grow larger and larger (Southwell et. al, 2010; Statistical Abstract of the United States 2013). Indeed, since 1970, the relationship between class size and students' ratings of teaching effectiveness has been the motivation of many studies. Such studies suggests a negative correlation between teaching effectiveness and class size (as class size increases STE ratings decrease) (Fourier, 2000; Mateo & Fernandez, 1996; Wood, et.al., 1974). Other research suggests this is because professors teaching small sections of students, are more able to increase student engagement, which can lead to higher retention of the material (Klegeris & Hurren, 2011) and to

subsequently higher STEs (Centra & Gaubatz, 2000). Interestingly, other research has found that attempting to moderate student engagement through interactive technology will not generate sufficient opportunities for intimate learning to offset the perception of decreased professor interaction (Blakemore, Switzer, DiLorio, & Fairchild, 1997). Conversely, other studies in the same period suggest this relationship between class size and STEs is spurious (Petchers & Chow, 1988; Lin, 1992, Aleamoni, & Graham, 1974; Dommeyer, 1997). Still others contradict these findings suggesting evidence of a curvilinear animosity effect; proposing students in larger and smaller classes may rate professors higher than those students in medium size classes. In this U-shaped relationship, researchers note that small (<25) and large classes (>50) have higher ratings, while medium (>26 & <60) have the lowest ratings (Fernandez, Mateo, & Muniz, 1999; Feldman, 1984). These somewhat conflicting results in conjunction with increasing college enrollment, and decreasing full time faculty make the topic relevant for further examination. To this end, the present research asserts that STEs in large classes be statistically different from those in small or medium classes. Said more formally:

H_2: Student teaching evaluations in large (≥ 50), medium (≥25 and <50) and small (<25) classes, will be statistically different.

1.3 Professor Gender

Mudding the waters of the logic of engagement and learning, is the impact of professor gender in the college classroom. Of particular interest to this research is the breath of studies suggesting STEs are affected by a student's previously established sex type roles (Feldman 1993; Statham, Richardson & Cook 1991; Basow & Silberg, 1987). Research has found, for example, that male students consistently rate female professors lower than their male colleagues, regardless of teaching style (Paludi & Bauer, 1983). A later study, focusing on student demographics and the evaluative process, suggests that female students are generally more discerning in evaluations and evaluate all courses, regardless of professor gender, more favorably than their male colleagues (Darby, 2006b). More directly relevant to this study is a matched pair study of award winning male and female professors. In this study students regularly rated the male professor higher than the female, and attributed the males' superior teaching skill to academic qualities; while attributing the female's skill to her attractiveness (Kaschak, 2006). Several years later in a comprehensive review of the literature Merritt (2008) found numerous studies that suggest STEs reflect student biases based on professor gender. Such studies date back to the 1970's and repeatedly report that gender, as well as a host of gender specific nonverbal actions (completely unrelated to teaching effectiveness) have been found to affect scores on STEs more than knowledge-based actions such as what the professor writes on the board. (Murray, 1997). Other earlier research calls into question the gender of the professor, suggesting gender of the professor does not matter to students when they are completing their STE (Elmore & LaPointe, 1974). In another study, researchers found that professors 'who looked distinguished and spoke authoritative' are rated more favorably than those without such characteristics (Backer, 2012). Similarly, other studies suggest good looking professors regardless of gender receive better rankings (Gray, & Bergmann, 2003). Given these contradictions in the literature, the following hypothesis is proposed:

H_3: Student teaching evaluations for male and female professors will be unequal.

1.4 Course Grades

The final topic of concern in this research is the impact of expected course grades on Student Teaching Evaluations. Historically, two schools of thought have been hypothesized. The first suggests a clear, positive relationship between expected course grades and STEs; and the second suggests the relationship between course grades and STEs is spurious and influenced by several other factors. Interestingly, early studies posit the only consistent finding between expected course grades and STEs is actually lack of consistency (Palmer, Carliner, & Romer, 1978; Kulik, & McKeachie, 1975). Yet, in a later meta-analysis of more than 41 studies, independently examining the relationship between course grades and STEs, found that indeed expected course grades were significantly and positively correlated with teaching effectiveness rating; providing some evidence for the influence of course grades and STEs (Cohen, 1981). Since this study, empirical research incorporating almost 40 years suggests STEs are higher when a student expects to earn a higher grade (Feldman, 1976, Goldberg & Callahan, 1991; Marsh & Rochi, 1997; Nasser & Hagtvet 2006). Yet, some controversy exists around the roll of student engagement. Indeed several studies suggest students who are more engaged are likely to both earn higher grades and rate professors more favorably (Culver, 2010). Still other researchers suggest that the correlation between course grades and STEs might be explained by cognitive dissonance theory; suggesting students expecting poor marks in a course will rate professors poorly to minimize ego threats (Maurer, 2006).

Of more concern however, is that in many academic settings STEs are a heavily weighted measure of teaching

effectiveness, and a major indicator in personnel decisions (Backer (2012). Such policies she notes encourages grade inflation, and leads to dumbing down of course material (Backer, 2012; Merritt, 2008; Wilson, 1998). Other studies find faculty indeed are subtlety encouraged to'*dumb down'* the material in an attempt to raise the overall course grades (Filetti, et. al., 2010, Centra, 2003, Wilson, 1998, Stratton, et. al., 1994). In several empirical studies of this relationship where thousands of faculty members were polled, 72% suggested they willingly *'watered down the material'* and inflated grades to enhance student evaluations (Backer, 2012; Gray & Bergmann, 2003). Making matters more complex are the empirical studies of students suggesting that they are increasingly aware of their in the faculty evaluation process and willingly admit acting in an opportunistic way to try to get better grades (Tucker, Jones, & Striker, 2008), and as many as 35% of those students polled in another study admitted that they would 'punish' their professors for failing their work by giving the professor low scores on the student evaluation (Backer, 2012).

Given this research it seems a variety of factors beyond student learning and teaching effectiveness may influence STE scores. In light of the controversial findings around this topic, it is important to examine the relationship between expected course grades and STEs. Specifically, the following hypothesis is proposed:

H_4: Student teaching evaluations and expected course grades will be positively correlated.

Finally, this research tests the interaction of the four main effects identified above. Given the controversy in the literature and between faculty members who are required to teach large sections of required classes, compared to those teaching large sections of elective classes, it seems relevant to test the combined effect on teaching evaluations of teaching a large required class, rather than a large elective class. Specifically, the following is hypothesized:

H_5: Student teaching evaluations in large required classes will be lower than those in large elective classes.

The sixth hypothesis examines the role of professor gender with respect to STEs. Specifically analyzing if male and female professors are equally able to effectively teach medium and large courses. If so, the STEs of male and female professors teaching medium and large sections of required and elective courses will not be statistically different. . As such, the following hypothesis is proposed:

H_6: The Student Teaching evaluations in medium and large sections of required classes will be statistically equal for male and female professors.

2. Methodology

2.1 Sample description

This study is designed to examine differences between several groups; therefore, it was necessary to collect a substantial sample of data, organize and clean the data, then partition it into several groups. Next it was necessary to insure the sample was independent, normally distributed and had equal population variances, such that an Analysis of Variance (ANOVA) could be utilized. Thus the present section is organized to insure the assumptions were met and teach each hypotheses. First a large set of independent STEs were obtained from a College of Business in a diverse Southwestern University will more than 7,000 students. The sample contains 30,571 evaluations obtained across six semesters between 2011 and 2013 for 255 professors. The final sample representing 1,057 classes ranging in size between 10 and 190 students. Each evaluation (class) is summarized as an average of ten teaching effectiveness measures. Next, all evaluations were coded to reflect the factors of interest in the study which are: class type (elective or required), class size (small, medium and large), professor gender and expected course grades.

Given the goal of this research was to analyze the effect of one noncontiguous factor on the category means of the STEs, Analysis of Variance (ANOVA) was chosen as the parsimonious method of choice for several hypotheses. In two of the hypotheses, (given the continuous nature of the independent variable), a linear regression model could be utilized in addition to the ANOVA.

Data analysis began with organizing and partitioning the data. First it was checked for normalcy using the Kolmogorov and the Levene Equal variance tests were conducted and neither were significant, suggesting the data is normal and has sufficiently similar variance to utilize the ANOVA method (table 1). Next, following previously established statistical methods; the difference between group means were tested were appropriate (Hair, Black, Babin, Anderson, & Tatham, 2006). The findings suggest that the data variance is not significantly different across group; thus appropriate for analysis.

Table 1. Levene Test of Homogeneity of Variance

Table 1	Levene Statistic	Sig.	Komogorov-Smirnov	Sig.
STEs	1.521	0.218	1,067	0.072

2.2 Analysis and Hypothesis testing

Using ANOVA, this research finds support for hypothesis 1 suggesting that STEs in elective classes will be higher than in required classes. ANOVA is the method of choice as it is efficient at determining if a statistically significant difference exists between the means of these two groups. Therefore the STE sample *(n =892)* was partitioned in to two groups; required courses (core business classes) and elective courses. Consistent with previous research (Johnson et. al, 2013) these data suggest scores obtained from STEs in elective courses were higher than those in required (or core) courses (table 2). Indeed, these analysis suggest in approximately 99% of the cases the STEs in elective classes will be statistically higher than those of required (core) classes. Specifically, using a 5 point scale, the scores obtained on STEs in elective courses will range from 4.24 to 4.30; while STEs in required (core) courses will range from 4.08 and 4.18. Thus the likelihood of two instructors (one teaching a required course the other teaching an elective course) obtaining similar STEs scores is less than 1% *(n=892; f = 9.89; p < .01)*. Said differently, such results make it unlikely that STEs obtained teaching a required course would be similar to those obtained teaching an elective.

Table 2. The effect of class type on STEs (ANOVA)

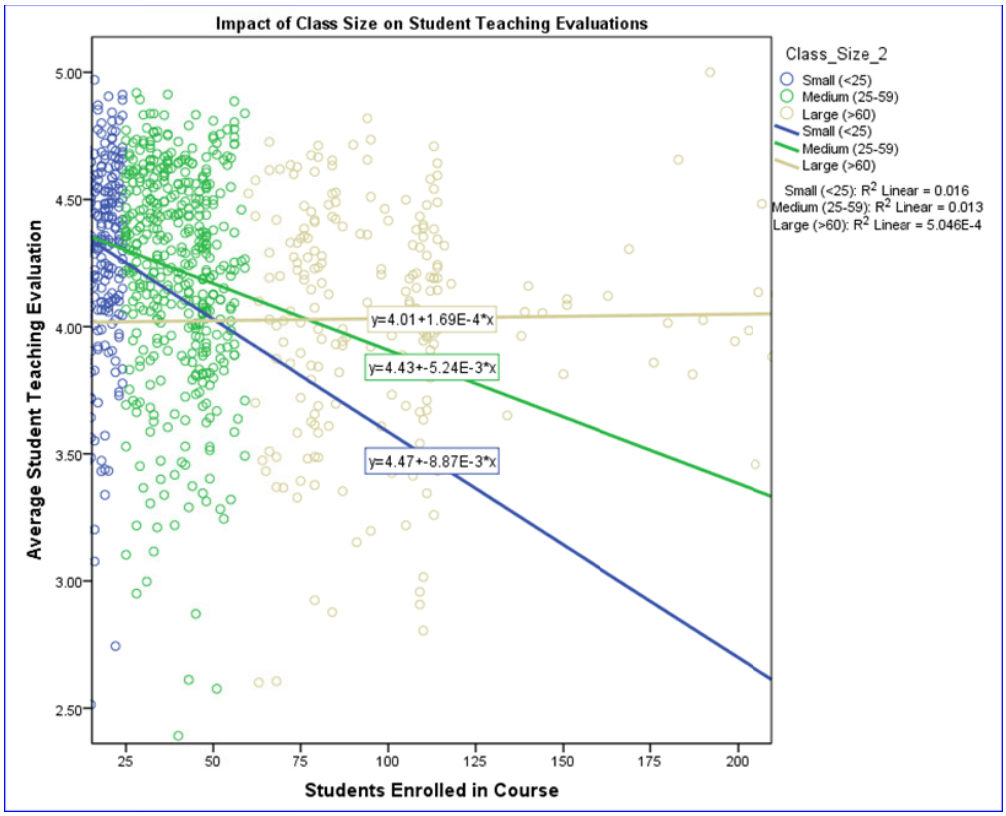
Class Type	*N*	*Mean*	σ	$\frac{\sigma}{\sqrt{n}}$	*Min*	*Max*	*F-Test*	*Sig*
Elective	522	4.253	0.417	0.023	4.24	4.30	9.89	0.01**
Required	370	4.162	0.440	0.018	4.08	4.18		
Overall	892	4.215	0.429	0.014	4.20	4.25		

***p < .01*

Hypothesis 2 is also supported suggesting scores obtained from STEs in large (≥ 50), medium (≥25 and <50) and small (<25) classes, are statistically different. Such a hypothesis is consistent with previous research suggesting STEs in large courses will be lower than those in smaller courses (Hippensteel, & Martin, 2005). Using ANOVA (is the method of choice, given its ability to discern differences between means) it was possible to determine if statistical differences exist between the STE scores obtained in small, medium and large classes differ. Indeed, the ANOVA suggested a statistically significant difference in the three groups was found *(n = 1,069; f = 39.843, p < .01)*. Next, given the nature of this hypothesis, post-hoc analysis was appropriate (Bonferroni). Such a procedure enables comparison of individual mean differences across all three class sizes, with respect to STEs. The results revealed that all (post-hoc) mean comparisons were statistically significantly different. Said differently, using a 5 point scale, classes with less than 25 students are likely to obtain STEs ranging from 4.30 to 4.38; while classes with between 26 and 50 students have STEs ranging from 4.18 to 4.26; and classes that have greater than 50 students are likely to have STEs ranging from 3.98 to 4.24 (table 3).

Table 3. The effect of class size on STEs (ANOVA and Post Hoc Analysis)

Class Size	*N*	*Mean*	*Σ*	$\frac{\sigma}{\sqrt{n}}$	*Lower Bound*	*Upper Bound*	*Class Size*	*Mean Difference*	*Sig.*
Small (<25)	382	4.341	0.405	0.020	4.300	4.382	Medium	0.119*	0.00**
							Large	0.306*	0.00**
Medium (25<50)	442	4.222	0.419	0.019	4.182	4.261	Small	-0.119*	0.00**
							Large	0.187*	0.00**
Large (>51)	249	4.035	0.440	0.028	3.979	4.241	Small	-0.306	0.00**
							Medium	-0.187*	0.00**



To graphically depict this relationship a regression model was run and graphed using class size groupings (as specified above) while leaving class size as a continuous variable. As can be seen in figure 1 the slope of the lines varies with class size, suggesting that class size impacts STE's.

Support for Hypothesis 3 was not found, suggesting that the STEs for male and females professors is not statistically significantly different. As in the previous hypotheses, an ANOVA was utilized to test this hypothesis. Specifically the data was reapportioned by gender *(n=1,053; male n = 774; female = 279*); and results interpreted. In this case, the analysis reveals that while the average STE scores of a male and female professors is different, they are not statistically different. Specifically, male professors average: 4.23 and female professors 4.20. Such a difference is indeed in the direction predicted, however not statistically so. Of note is that the scores obtained (again on a 5 point scales) by males range from 4.96 and 4.26; while those of his female colleagues are likely to range between 4.15 and 4.25. The overlap of 0.02 makes it impossible to conclude that the STEs differ based on gender. Thus, we do not find support for hypothesis 3.

Table 4. The effect of gender on STEs (ANOVA)

Gender	*N*	*Mean*	σ	$\frac{\sigma}{\sqrt{n}}$	*Lower Bound*	*Upper Bound*	*F*	*Sig.*
Male	774	4.227	.433	.016	4.196	4.258	0.934	0.334
Female	279	4.198	.442	.026	4.146	4.250		
Total	1,053	4.219	.435	.013	4.193	4.246		

Given the continuous nature of expected course grade, a linear regression model was used to test and find support for hypothesis 4; suggesting that indeed STEs are positively correlated to expected course grade. Of specific interest to this research is the relationship between the outcome variable, *STE* (and the independent variable *expected course grade* (B_1). As can be seen in table 5 and in figure 1, the regression model supports this hypothesis, suggesting *STE* increases in a linear manner with respect to expected course. Specifically, this analysis suggests that for each 1-unit increase in expected course grade, *STEs* are likely to increase by about *0.37 (n = 836; f = 169.762, p < 0.01)*. Further support for this model can be found in the coefficient of determination (R^2). This goodness of fit measure, suggests that about 14% of the variation in *STEs* can be explained by *expected course grade)* a statistically significant finding (table 5). This relationship is also graphically depicted in figure 1.

Table 5. The effect of course GPA on STEs (Regression)

Course GPA	N	F	Sig	Standardized Coefficient	*t-Statistic*	Sig
Undergraduate GPA	836	169.762	0.000**	.370	13.02	0.01**

*** p < .01; Independent Variable: Class GPA; Dependent Variable: STEs*

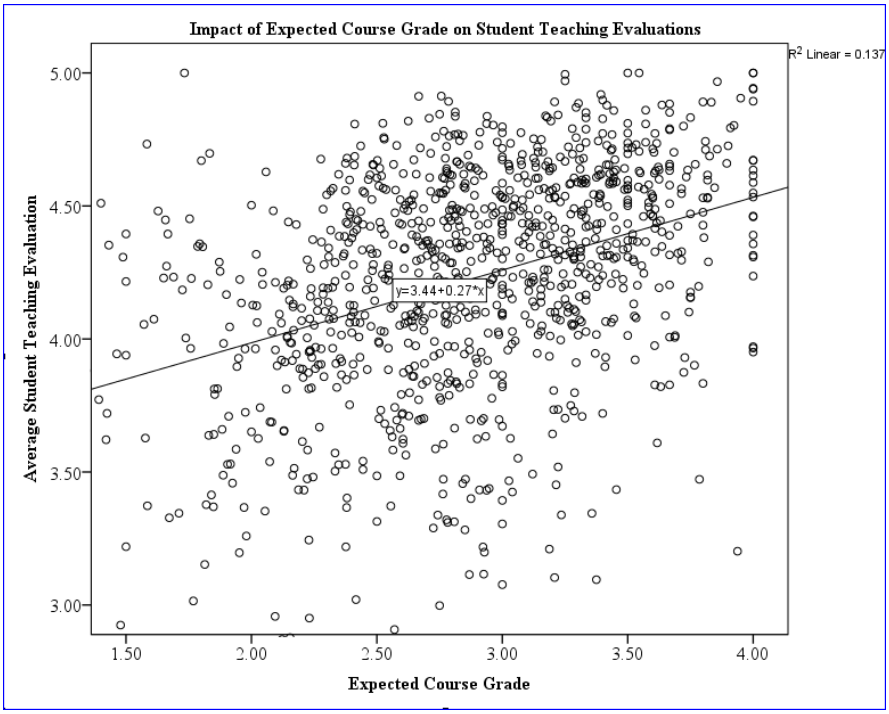


Figure 1. Impact of Expected Course Grade on Student Teaching Evaluations

Support was also found for Hypothesis 5 suggesting that STEs in large required classes are statistically lower than the STEs of large elective classes. This hypothesis required the creation of two groups of data: group 1 representing STEs in large (*n >50*) required classes and a second group consisting of large elective classes (*n>50*). Once partitioned the ANOVA *(n=239; f=7.97; p<.01**)* suggests differences in the group means exists; specifically large required classes have STEs ranging from 3.91 to 4.03, while large elective classes have STEs ranging from 4.04 to 4.25 (both on a 5 point scale); a statistically significant difference (table 6).

Table 6. The effect of class type and class size on STEs (ANOVA)

Class type and Class Size	*N*	*Mean*	*σ*	$\frac{\sigma}{\sqrt{n}}$	*Min*	*Max*	*Mean Square*	*F*	*Sig.*
Large Electives	73	4.15	0.45	0.05	4.04	4.25	1.50	7.97	0.01**
Large Required	166	3.97	0.43	0.03	3.91	4.03	0.19		
Total	239	4.03	0.44	0.03	3.97	4.08			

*** p < .01; Independent Variable: class size; Dependent Variable: STEs*

The final hypothesis asserts STEs for male and female professors teaching large sections of required classes will be equal. Again following the previous pattern the data was portioned into two groups; group 1 consisting of large required classes taught by males; and group 2 large required classes taught by females. Once portioned the ANOVA suggests there is a statistical difference between the group means. Indeed, *(n= 166; f = 21.19; p < .01**)*, these data support hypothesis 6. That is, on average large required classes taught by male professors are likely to have STE's ranging from 3.97-4.11; while large required classes taught by female professors are likely to have STEs ranging from 3.51 to 3.83. Of note, it appears that a female professor is not likely to score in the "4" range (while teaching a large section of a required course) while her male colleague is likely to do so (table 7).

Table 7. The interaction effect of class size, class type and gender on STEs

STEs	*N*	*Mean*	σ	$\frac{\sigma}{\sqrt{n}}$	*Min*	*Max*	*F*	*Sig*
Large Required (Male)	136	4.05	0.40	0.03	3.97	4.11	21.19	0.01**
Large Required (Female)	30	3.67	0.43	0.08	3.51	3.83		
Total	166	3.98	0.43	0.03	3.91	4.04		

** *Significant p < 0.01*

3. Discussion

This research provides interesting insight into several important factors that seem to consistently affect Student Teaching Evaluations (STEs) and are beyond the control of the professor being evaluated. As can be seen in the previous pages scores on STEs for those teaching required classes are not likely to be consistent with those teaching electives. Likewise for those teaching large classes, again they are less likely to receive STE scores that are statistically similar to their colleagues teaching smaller classes. Also surprising were the results with respect to gender. Women seem to be competitive with their male colleagues on many levels, but when it comes to teaching large classes their scores on STEs seem to drop substantially. Such findings should be examined across additional institutions. While such findings are a good reference point for women teaching large sections of courses, others evaluating women should note this as well. Further research may be warranted to determine why this drop occurs when women teach large sections of courses and what can be done to mediate this situation. In a similar vein some attention should be paid to course grades and STEs. Given the presence of this positive relationship it seems to suggest that another measure of student learning may be in order. Such findings certainly call into question the assertion that (because of its numerical nature) STE data is 'objective', and further generate some reflection about the wide spread use and reliability of this instrument.

In these times of increased curriculum rigor, limited resources and increasing class sizes, this presents a difficult problem with deep impact. The results of this study coupled with the lack of pedagogical training provided to most professors, the quickly changing classroom management systems, and not to mention on-line homeworkbring to light a daunting problem. At a minimum the present research calls into question policies and procedures that have been standard for many years. The practice of practice of taking STEs at face value should be reconsidered. While this research is not comprehensive, it certainly provides the necessary background to ask those currently using these measures as the sole indicator of teaching effectiveness to at least reconsider their systems. Utilizing information such as this to influence personnel decisions should be examined, and used at least in conjunction with other measures of teaching effectiveness.

Table 8. Short Summary of hypothesis

	Hypotheses results summary	Results	Statistical Summary	Significance
1	The effect of class type on STE (ANOVA*)*	Supported	Elective: 4.24 – 4.30 Required: 4.08 – 4.18	*n = 1,069* *f = 23.72, p <0.01***
2	The effect of class size on STEs (ANOVA & Post Hoc)	Supported	Small: 4.17 – 4.32 Medium: 4.13-4.21 Large: 3.92 – 4.07	*n = 838* *f = 18.68 p <0.01***
3	STEs for males will be different that STEs for females	Not Supported	Males: 4.19 – 4.28 Females: 4.14 – 4.26	*N=1,053* *f = 0.934, p <0.334*
4	STEs will be positively correlated with course GPA	Supported	*β =0.37* $R^2 = 0.14$	*n = 836* *f = 105.17, p<0.01***
5	STEs of Large Required Classes will be < than STEs of Large Elective Classes	Supported	Large Electives: 4.04-4.25 Large Required: 3.91-4.03	*n=239* *f = 7.97, p <0.01***
6	STEs of female taught large required classes < STEs of male taught large required classes	Supported	Large RQ Female: 3.51-3.83 Large RQ Male: 3.97 – 4.11	*n =166* *f = 21.19, p <0.01***

4.1 Limitations

Certainly, caution should be taken when interpreting the results of this study. While every effort has been made to insure reliability and validity of these constructs, it is always necessary to exercise thoughtfulness with generalizing the results to universities that perhaps differ in terms of demographic composition, class size, funding, and subjects. The data may suggest some broad trends; however, it is best to conduct similar analysis at individual universities.

References

Aleamoni, L. & Graham, M. (1974). The Relationship Between CEQ Ratings and Instructors Rank, Class Size and Course Level. *Journal of Educational Measurement*, 11(3), 189-191. http://dx.doi.org/10.1111/j.1745-3984.1974.tb00990.x

Backer, E. (2012). Burnt at the Student Evaluation Stake – the Penalty for Failing Students. E-*Journal of Business Education & Scholarship of Teaching*, 6(1) 1-13.

Basow , S. & Silberg, N. (1987). Student Evaluations of College Professors: Are Female and Male Professors Rated Differently? *Journal of Educational Psychology*, 79(3), 308-314. http://dx.doi.org/10.1037/0022-0663.79.3.308

Blakemore, O., Switzer, J., DiLorio, J., & Fairchild, D. (1997). Exploring the Campus Climate for Women Faculty. In Benokraitis, N.V. (Ed.), Subtle Sexism: Current Practice and Prospects for Change, 54 – 71. Thousand Oaks, CA: Sage.

Bosshardt, W. & Watts, M. (2001). Comparing Student and Instructor Evaluations of Teaching. *The Journal of Economic Education*, 32(1), 3-17. http://dx.doi.org/10.1080/00220480109595166

Bray, J., & Howard, G. (1980). Interaction of Teacher and Student Sex and Sex Role Orientations and Student Evaluations of College Instruction. *Contemporary Educational Psychology*, 5(3), 241-248. http://dx.doi.org/10.1016/0361-476X(80)90047-8

Centra, J. (2003). Will Teachers Receive Higher Student Evaluations by Giving Higher Grades and Less Course Work? *Research in Higher Education*, 44(5), 495-519. http://dx.doi.org/10.1023/A:1025492407752

Centra, J., & Gaubatz N. (2000). Is there Gender Bias in Student Evaluations of Teaching? *The Journal of Higher Education,* 71(1), 17-33. http://dx.doi.org/10.2307/2649280

Cohen, P. (1981). Student Rating of Instruction and Student Achievement: A Meta-Analysis of Multisession Validity Studies. *Review of Educational Research*, 51(3), 281-309. http://dx.doi.org/10.3102/00346543051003281

Costin, F., Greenough, W. T., & Menges, R. J. (1971) Student Ratings of College Teaching: Reliability, Validity, and Usefulness. *Review of Educational Research*, 41(5), 511-535. http://dx.doi.org/10.3102/00346543041005511

Culver, S. (2010). Course Grades, Quality of Student Engagement, and Students' Evaluation of Instructor. *International Journal of Teaching and Learning in Higher Education,* 22(3), 331 – 336.

Darby, J. (2006a) The Effects of Elective or Required Status of Courses on Student Evaluations. *Journal of Vocational Education & Training*, 58(1), 19-29. http://dx.doi.org/10.1080/13636820500507708

Darby, J. (2006b) Evaluating Courses: An Examination of the Impact of Student Gender. *Educational Studies*, 32(2), 187-199. http://dx.doi.org/10.1080/03055690600631093

Dommeyer, C.J. (1997). Class Size in An Introductory Marketing Course: Student Attitudes, Evaluations, and Performance. *Marketing Education Review*, 7(1), 13 – 25.

Drago, W., & Wagner, R. (2004). Vark Preferred Learning Styles and Online Education, *Management Research News*, 27(7), 1–13. http://dx.doi.org/10.1108/01409170410784211

Eisenhardt, K. (1989) Agency Theory: An Assessment and Review, *Academy of Management Review,* 14(1), 57-74. http://dx.doi.org/10.2307/258191

Eiszler, C. (2002). College Student Evaluations of Teaching and Grade Inflation. *Research in Higher Education*, 43(4), 483–501. http://dx.doi.org/10.1023/A:1015579817194

Elmore, P., & LaPointe, K. (1974). Effects of Teacher Sex and Student Sex on the Evaluation of College Instructors. *Journal of Educational Psychology*, 66(3), 386-389. http://dx.doi.org/10.1037/h0036493

Feldman, K. (1976). Grades and College Students' Evaluations of Their Courses and Teachers. *Research in Higher Education*, 4(1), 69 – 111. http://dx.doi.org/10.1007/BF00991462

Feldman, K. (1984). Class Size and College Students' Evaluations of Teachers and Courses: A Closer Look. *Research in Higher Education,* 21(1), 45-116. http://dx.doi.org/10.1007/BF00975035

Feldman, K. (1993). College Students' Views of Male and Female College Teachers: Part II-Evidence from Students' Evaluations of Their Classroom Teachers. *Research in Higher Education*, 34, 151-211. http://dx.doi.org/10.1007/BF00992161

Fernandez, J., Mateo, M., & Muniz, J. (1999). Is There a Relationship Between Class size and Student Ratings of Teaching Quality? *Educational and Psychological Measurement*, 58(4), 596-604. http://dx.doi.org/10.1177/0013164498058004003

Filetti, J., Wright, M., & King K. (2010) Grades and Ranking: When Tenure Affects Assessment. *Practical Assessment, Research & Evaluation,* 15(14), 2-5.

Fourier, C. (2000). Class Size and Student Ratings of Teaching Presentation. *South African Journal of Higher Education,* 14(3), 132-138.

Goldberg, G. & Callahan, J. (1991). Objectivity of Student Evaluations of Instructors. *Journal of Education for Business,* 66 (6), 377-379. http://dx.doi.org/10.1080/08832323.1991.10117505

Gomez-Mejia, L. & Balkin, D. (1992) Determinants of Faculty Pay: an Agency Theory Perspective. *Academy of Management Journal,* 35(5), 921-955. http://dx.doi.org/10.2307/256535

Gray, M., & Bergmann, B. (2003). Student Teaching Evaluations: Inaccurate, Demeaning, Misused. *Academe Online,* 89(5), 44-46. http://dx.doi.org/10.2307/40253388

Hair, J., Black, W., Babin, B., Anderson, R., & Tatham, R. (2006). *Multivariate Data Analysis.* New Jersey, Pearson-Prentice Hall.

Hippensteel, S.P. & Martin, W. (2005). Increasing the Significance of Course Evaluations in Large-Enrollment Geoscience Courses. *Journal of Geoscience Education,* 53(2), 158 – 165.

Houston, D. Meyer, L., & Paewai, S. (2006) Academic Staff Workloads and Job Satisfaction: Expectations and values in academics. *Journal of Higher Education Policy and Management,* 28(1), 17-30. http://dx.doi.org/10.1080/13600800500283734

Johnson, M.D., Narayanan, A., & Sawaya, W.J. (2013). Effects of Course and Instructor Characteristics on Student Evaluation of Teaching across a College of Engineering. *Journal of Engineering Education,* 102(2), 289 – 318. http://dx.doi.org/10.1002/jee.20013

Kaschak, E. (1981). Another Look at Sex Bias in Students' Evaluations of Professors: Do Winners Get the Recognition That They Have Been Given? *Psychology of Women Quarterly*, 5(5), 767-772. http://dx.doi.org/10.1111/j.1471-6402.1981.tb01100.x

Klegeris, A. & Hurren, H. (2011). Impact of Problem-Based Learning in a Large Classroom Setting: Student Perception and Problem-Solving Skills. *Advances in Physiology Education,* 35(4), 408 – 415. http://dx.doi.org/10.1152/advan.00046.2011

Knol, M., Veld, H. & Mellenbergh, G. (2013) Experimental Effects of Student Evaluations Coupled with Collaborate Consultation on College Professors' Instructional Skills. *Research in Higher Education*, 54, 825-850. http://dx.doi.org/10.1007/s11162-013-9298-3

Kulik, J. & McKeachie, W. (1975). The Evaluation of Teachers in Higher Education. *Review of Research in Education*, 3(2), 210-240. http://dx.doi.org/10.2307/1167259

Kuo, W. (2007). Editorial: How Reliable is Teaching Evaluation? The Relationship of Class Size to Teaching Evaluation Scores. *IEEE Transactions on Reliability,* 56(2), 178 – 181. http://dx.doi.org/10.1109/TR.2006.874909

Lin, W. (1992). Is Class Size a Bias to Student Evaluations of Faculty? A Review. *Chinese University Education Journal,* 20(1), 49-53.

Lovell, G. & Haner, C. (1955). Forced Choice Applied to College Faculty Rating, *Educational and Psychological Measurement*, 15(1), 291–304. http://dx.doi.org/10.1177/001316445501500309

Marsh, H. (1984). Students' Evaluation of University Teaching: Dimensionality, Reliability, Validity, Potential Biases, and Utility. *Journal of Educational Psychology*, 76(5), 707-754. http://dx.doi.org/10.1037/0022-0663.76.5.707

Marsh, H., & Roche, L. (1997). Making Students' Evaluations of Teaching Effectiveness Effective: The Critical Issues of Validity, Bias, and Utility. *American Psychologist*, 52(11), 1187. http://dx.doi.org/10.1037/0003-066X.52.11.1187

Mateo, A., & Fernandez, J. (1996). Incidence of Class Size on the Evaluation of University Teaching Quality. *Educational and Psychological Measurement,* 56(5), 771-778. http://dx.doi.org/10.1177/0013164496056005004

Maurer, T. (2006). Cognitive Dissonance or Revenge? Student Grades and Course Evaluations, *Teaching of Psychology*, 33(3), 176-179. http://dx.doi.org/10.1207/s15328023top3303_4

McKeachie, W. (1990). Research on College Teaching: The Historical Background. *Journal of Educational Psychology*, 82(2), 189-200. http://dx.doi.org/10.1037/0022-0663.82.2.189

Merritt, D. J. M. (2008). Bias, the Brain, and Student Evaluations of Teaching. *St. John's Law Review*, 28, 235-287.

Murray, H.G. (1997). Effective Teaching Behaviors in the College Classroom. In: Effective Teaching in Higher Education: Research and Practice, Perry, R.P. & Smart, J.C (eds). Agathon Press: Bronx, NY.

Nasser, F. & Hagtvet, K. (2006). Multilevel Analysis of the Effects of Student and Instructor/Course Characteristics on Student Ratings. *Research in Higher Education*, 47(5), 559-590. http://dx.doi.org/10.1007/s11162-005-9007-y

Ory, J. (2001). Faculty Thoughts and Concerns about Student Ratings. *New Directions for Teaching & Learning*, 87, 3–15. http://dx.doi.org/10.1002/tl.23

Palmer, J., Carliner, G., & Romer, T. (1978). Leniency, Learning, and Evaluations. *Journal of Educational Psychology*, 70(5), 855-863. http://dx.doi.org/10.1037/0022-0663.70.5.855

Paludi, M., & Bauer, W. (1983). Goldberg Revisited: What's in an Author's Name. *Sex Roles*, 9(3), 387-390. http://dx.doi.org/10.1007/BF00289673

Petchers, K., & Chow, C. (1988). Sources of Variation in Students' Evaluations of Instruction in a Graduate Social Work Program. *Journal of Social Work Education,* 24(1), 35-42.

Pohlmann, J. (1975). A Multivariate Analysis of Selected Class Characteristics and Student Ratings of Instruction. *Multivariate Behavioral Research*, 10(1), 81–91. http://dx.doi.org/10.1207/s15327906mbr1001_5

Southwell, D., Gannaway, D., Orrell, J., Chalmers, D., Abraham, C. (2010). Strategies for Effective Dissemination of the Outcomes of Teaching and Learning Projects. *Journal of Higher Education Policy & Management*, 32(1) 55-67. http://dx.doi.org/10.1080/13600800903440550

Statham, A., Richardson, L., & Cook, J. (1991). Gender and University Teaching: A Negotiated Difference. Albany, NY: State University of New York Press.

Statistical Abstract of the United States, Table 296: *Employees in Higher Education Institutions,* 1995 to 2013.

Stratton, R., Myers S., & King, R. (1994). Faculty Behavior, Grades, and Student Evaluations. *The Journal of Economic Education*, 25(1), 5-15. http://dx.doi.org/10.1080/00220485.1994.10844810

Tucker, B., Jones, S., & Straker, L. (2008). Online Student Evaluation Improves Course Experience Questionnaire Results in a Physiotherapy Program. *Higher Education Research & Development*, 27(3), 281-296. http://dx.doi.org/10.1080/07294360802259067

Whitten, B., & Umble, M. (1980). The Relationship of Class Size, Class Level, & Core vs. Non-Core Classification for a Class to Student Ratings of Faculty: Implications for Validity, *Educational and Psychological Measurement,* 40(2), 419 – 423. http://dx.doi.org/10.1177/001316448004000220

Wilson, R. (1998). New Research Casts Doubt on Value of Student Evaluations of Professors, *The Chronicle of Higher Education,* 16, A12-A14.

Wood, K., Linsky, A., & Straus, M. (1974). Class Size and Student Evaluations of Faculty, *The Journal of Higher Education,* 45(7), 524-534. http://dx.doi.org/10.2307/1980791