Effects of Course and Instructor Characteristics on Student Evaluation of Teaching across a College of Engineering

Michael D. Johnson,[a] Arunachalam Narayanan,[b] and William J. Sawaya[c]

[a]*Texas A&M University,* [b]*University of Houston,* [c]*Bowling Green State University*

Abstract

Background Student evaluations of teaching (SETs) are a widely used metric to evaluate instructor effectiveness and are used to make promotion, tenure, and retention decisions for faculty. There is also growing interest by those outside the university community to use these metrics to evaluate faculty and broader academic performance.

Purpose (Hypothesis) This study seeks to understand if and how course and instructor characteristics affect SETs and thereby to improve the usefulness of these metrics. This article aims to statistically examine the relationship between course and instructor characteristics and SETs.

Design/Method SETs from a large engineering college at a major public university were evaluated over a seven-semester period that covered 3938 courses taught by 549 unique engineering instructors. Course and instructor demographic data were statistically evaluated for their effects on SETs.

Results Course characteristics such as class size, course level, and whether a course was an elective or required had statistically significant effects on SETs. Instructor characteristics of gender and academic rank affected SETs and average course grades, respectively. Average course grades were positively correlated with SETs.

Conclusions Data analysis showed that course characteristics, faculty demographics, and average course grades had statistically significant effects on SETs; however, in some cases the effect sizes of these variables were small. Administrators and senior faculty members should be cognizant of these relevant factors and their effects when assigning faculty to certain courses and evaluating their teaching effectiveness using SETs.

Keywords student evaluations of teaching (SETs); academic administration; engineering

Introduction

University faculty, administrators, students, and often parents have a vital interest in the quality of instruction provided within universities. Student evaluations of teaching (SETs) are widely used to evaluate the instructional performance of faculty and for evaluating them for promotion, pay, tenure, and hiring decisions (Langbein, 2008). While it is likely that the use of SET data will increase as accountability in higher education is demanded, some faculty are hostile toward the process of evaluation (Feldman, 2007). Indeed, simply comparing

Journal of Engineering Education
April 2013, Vol. 102, No. 2, pp. 289–318

http://wileyonlinelibrary.com/journal/jee
DOI 10.1002/jee.20013

scores on their face value might not give a complete assessment of teaching effectiveness. Course or instructor characteristics may significantly affect the SETs. As SETs become increasingly available for administrators and other university stakeholders to use, a better understanding of how they may be influenced by course grades, faculty demographics, and course characteristics is critical so that instructors are not unduly criticized (or praised), when in fact the SETs may be highly influenced by factors over which faculty have little control. Previous work has shown that SETs should not be used to compare faculty without considering additional factors (Marsh, 1984; Arreola, 2007; Feldman, 2007; Zabaleta, 2007). Therefore, understanding how course and demographic factors may affect SETs can guide administrators about courses assignments so that the instructors have the best opportunity to succeed and that the students have the greatest opportunity to learn.

Understanding the perceptions of faculty members towards teaching evaluations over time is key in understanding and interpreting the importance of SETs. In a 1989 survey of faculty, over two-thirds of faculty reported that SETs were important for granting tenure at their institution (Boyer, 1990). In that study the engineering faculty subgroup comprised 55%, and faculty members younger than 40 years of age considered teaching to be more important than the overall faculty respondents. Today these faculty members are among the senior faculty and administrators at colleges of engineering in the United States. Despite their wide use, the reliability and validity of SETs is debated. Marsh (1984) made a detailed case for the SET and noted that peer ratings (another often-proposed way to evaluate teaching effectiveness) are not well correlated with each other or other measures of teaching effectiveness. Cohen (1981) found that SETs are highly correlated objective measures of student achievement (e.g., common final exams). On the other hand, Johnson (2003) noted that SETs are not reliable for assessing teaching effectiveness, cited them as a cause for grade inflation, and decried their ubiquity in personnel assessment decisions. But Franklin, Theall, and Ludlow (1991) found no evidence that SETs contributed to grade inflation.

Kapel (1974) noted that for a teaching evaluation tool to be useful, it "should be valid, reliable, easy to administer, easy to fill out, applicable for many teaching styles and conditions, easy to analyze, and capable of providing meaningful information for the sundry publics interested in the results." Divoky and Rothermel (1988) stated a valid instrument should assess five dimensions of teaching: (1) delivery – the instructor's ability and way of conveying material; (2) depth of knowledge – the instructor's mastery of the subject matter; (3) interpersonal skills – the way in which the instructor interacts with students on a personal and/or professional level; (4) organization – the instructor's arrangement of each lecture and of the course material in general; and (5) relevance – the instructor's ability to relate the subject matter to what is important and meaningful to students.

Because factors that are not related to student learning (but that may affect SETs) are central in discussions of the overall validity of SETs for their purposes, this article examines SETs in a college of engineering in a large public university over seven semesters. The purpose of this study is not to examine the relationship between SETs and teaching effectiveness, but to analyze the role of course grades, instructor demographics, and course characteristics on students' evaluations of teaching. If such variables play a significant role in determining an instructor's performance according to an SET, then administrators should take that role into account (Pohlmann, 1975; Blackhart, Peruche, DeWall, & Joiner, 2006); in some cases they might need to control for biases in the evaluation system (Cashin, 1990). While numerous studies have examined large datasets from fields such as economics (McPherson, Jewell, & Kim, 2009) or language arts (Zabaleta, 2007), this study

presents one of the few large-scale analyses of engineering SETs. McPherson and Jewell (2007) noted that one could use data to determine an expected evaluation score for a given faculty member based on his or her individual and course characteristics; an individual faculty member could then be compared to this baseline. This would make identifying which characteristics (not related to student learning) affect SETs a crucial part of the instructor evaluation process.

Context and Hypotheses

SETs are used to evaluate faculty in varied and complex contexts. The broad classes of course and instructor characteristics considered here are those related to the courses themselves and those specific to instructors. Both types of characteristics are also examined relative to the grades assigned by instructors in developing the following hypotheses.

Course Characteristics

Clayson (2009) noted that if the SET process is valid, students' ratings of their teachers performance should be correlated with student learning. Exogenous factors unrelated to instructor performance should not affect the aggregate SET score (henceforth SET scores). There is, however, evidence that certain class-related factors affect student evaluations, all other things being equal (e.g., Zabaleta, 2007).

Class size One of the more researched class-related factors that may affect SETs is class size. We define class size as the number of students participating in the main lecture portion of a course. Class size is important because faculty generally have little control over their class size; and if significant, the effects of class size should be included in interpretations of evaluations. Research on the effect of class size on the SET is equivocal. Lin (1992) reviewed several studies and determined that class size does not bias the SET. No statistically significant relationship for class size and SET was found in a large sample of language courses (Zabaleta, 2007). Marsh and Roche (1997) also reviewed several previous studies but found that in some cases class size can significantly affect the SET, but in other cases the effects were uncertain.

Significantly more evidence points towards a negative effect of class size on SETs. Watkins (1990) analyzed a large dataset of 20,000 ratings from over 200 courses and found class size had a significant negative correlation with SET scores. A large study of economics students found that while class size had a negative effect on SET scores in lower-level (or principles) courses, it had a slightly positive effect in higher-level courses (McPherson et al., 2009). When class size is treated as a categorical value (e.g., large, medium, small), larger courses have lower SETs (Lovell & Haner, 1955; Hippensteel & Martin, 2005; Ragan & Walia, 2010). A significant negative correlation occurs when class size and SET are compared directly (Franklin et al., 1991; Ellis, Burke, Lomire, & McCormack, 2003; Burns & Ludlow, 2005; Westerlund, 2008; Chapman & Ludlow, 2010; Ragan & Walia, 2010).

In some cases, the relationship between class size and SET has been found to be neither linear nor consistent. Where a nonlinear relationship has been shown, SET scores decrease with increasing class size to some point, and then begin to increase again (Wood, Linsky, & Straus, 1974; Marsh, Overall, & Kesler, 1979; Mateo & Fernandez, 1996; El Ansari & Oskrochi, 2006).

The following hypotheses are proposed in light of these previous findings:

H_{1a} There is a negative linear correlation between class size and SET scores.

H_{1b} There is a nonlinear parabolic relationship between class size and SET scores.

H_{1c} Upper-level courses have lower negative correlations between class size and SET scores than do lower-level courses.

Course level Course level is another factor that has been researched extensively. Students may adapt their approach towards the SET as they progress in their academic careers and mature. Course level is typically viewed as a proxy for student maturity. A direct assessment, however, of student interests found that older students were more interested in instructors who are dedicated and motivating, while younger students were less interested in how adaptive and knowledgeable the instructor was (Donaldson, Flannery, & Ross-Gordon, 1993). Few studies evaluating SETs have examined courses across varying levels, and among these, findings are inconclusive. Several investigators found little or no effect of course level on the SET (Scherr & Scherr, 1990; Troy, Friedrich, & Troy, 1992; Blackhart et al., 2006; Morgan & Davies, 2006). Those that did show a relationship, found this correlation to be small: 6% lower for lower-level courses (Hippensteel & Martin, 2005); or in some cases a small correlation that was not significant (Ellis et al., 2003). Peterson, Berenson, Misra, and Radosevich (2008) focused on 400-level courses and found that those students rated their instructors higher than all other courses evaluated.

In addition to a direct examination of the course level and SET, the effect of course level on relationships between other variables and the SET has also been examined. McPherson (2006) found that while class size and instructor experience have no effect on SETs for upper-level courses, there is an effect for lower-level courses. Bailey, Gupta, and Schrader (2000) observed that lower-level students expect the instructor to be more involved and interactive. Sailor, Worthen, and Shin (1997) found positive correlations between grades and SETs for undergraduate classes, but it was negative for graduate classes. There was no difference between upper- and lower-level undergraduate courses with respect to the positive correlation. In Feldman's (1993) review, conflicting data regarding the role of gender on SETs by course level was cited. In one case, male instructors scored higher in upper-level courses; in the other, they scored lower.

This prior work led to the following hypotheses:

H_{2a} There is a positive linear correlation between course level and SET scores.

H_{2b} There is a positive correlation between instructor experience and SET scores that is higher in lower-level courses than upper-level courses.

H_{2c} There is a difference between upper- and lower-level courses with respect to the effect of instructor gender on SET.

Course type In some departments (e.g., math), the majority of courses taught are service courses taken by students not majoring in that subject, while in other departments (e.g., civil engineering), the vast majority of courses are for majors. In addition to this core (a required math course for an engineering major) and major (a required civil engineering course) demarcation, there are also electives. Gage (1961) posited that students taking a course as an elective may have more interest in the subject and therefore might assign higher SET ratings to it. Lovell and Haner (1955) had previously found that mean SET scores were lower for required than for elective courses. Kapel (1974) also found that students taking electives and graduate students assign more higher SET scores; Aleamoni and Thomas (1980) also found

elective and minor courses received higher SET scores. McPherson (2006) used the percentage of major students and found a significant negative correlation between this variable and SET score. Peterson et al. (2008) found that students in core courses rated faculty lower than those in either elective or major courses. Pohlmann (1975) found that a higher percentage of students taking a course as an elective led to a higher SET score.

In cases where more detailed data have been examined, some factors are more important with respect to major or required courses than others. Divoky and Rothermel (1988) broke courses into four categories: non-major required, non-major elective, major required, and major elective. They then compared the importance of the previously noted dimensions of teaching (delivery, depth of knowledge, interpersonal skills, organization, and relevancy) for these courses. They found delivery to be more important for non-major required courses; depth of knowledge more important for major electives; and interpersonal skills for major required courses. In Adams' (2005) study of student and faculty perceptions of which factors should influence course grades, there was a large discrepancy between what faculty and students believed should influence grades for elective and liberal arts courses. Students believed that effort should have a much greater influence on grades, while faculty placed more emphasis on achievement. The effect of course type and the correlation between grades and SET scores has also been explicitly analyzed; a higher correlation was found between grades and SET score for required than for elective courses (DuCette & Kenney, 1982).

In the context of previous work, the following hypotheses are proposed:

H_{3a} Elective courses have higher SET scores than major required and general required courses.

H_{3b} The positive correlation between average course grades (henceforth course grades) and SET scores is higher in major required and general required courses than in elective courses.

Instructor Demographics

Ideally, the demographic profile of instructors would be immaterial to their teaching effectiveness. Unfortunately, this is often not the case, as shown by many studies that have evaluated the influence of instructor characteristics on SETs. While personality traits such as warmth or charisma (Bennett, 1982) and ethnicity (McPherson & Jewell, 2007) have been evaluated elsewhere, this article focuses on three objective demographic characteristics: gender, experience, and academic rank. The identification of demographic biases in SETs can allow for them to be controlled for, allowing SETs to evaluate the performance of instructors rather than their demographic characteristics (Dresel & Rindermann, 2011).

Gender The interaction of gender and SET has been evaluated for both the student and the instructor. While the student gender composition of a class would be deemed a course characteristic, the interaction of student and instructor gender is often presented together. Aleamoni and Thomas (1980) reported that female students gave higher ratings; this finding conflicts with Kapel's (1974) report that male students gave higher SETs. Morgan and Davies (2006) analyzed an all-female student population and found that male instructors had higher SETs than their female counterparts. McPherson et al. (2009) also reported that male instructors received higher SETs in a mixed-gender student population.

Given the numerous factors that influence SETs, the many studies that have evaluated the role of instructor gender have produced conflicting results. Several evaluations found that there was no or only an insignificant effect of instructor gender on SETs (Feldman,

1993; Centra & Gaubatz, 2000; Ellis et al., 2003; Blackhart et al., 2006; McPherson & Jewell, 2007; Zabaleta, 2007). By focusing on aggregate (a combined score using all SET questions) SET, these results may not capture the way that students are assessing instructors of differing genders. Bennett (1982) reported that female instructors were expected to provide more time to students and received higher ratings on interpersonal aspects of teaching. Statham, Richardson, and Cook (1991) noted that female instructors who interacted more with students received higher competency ratings. Clayson (2009) also found that female instructors were expected to be more nurturing than their male counterparts. Female instructors who self-reported themselves as "warm" received higher effectiveness ratings than their male colleagues who also self-reported themselves as "warm" (Elmore & LaPointe, 1975).

Given this context, we propose the following hypotheses:

H_{4a} There is a higher aggregate SET score for male instructors than for female instructors.

H_{4b} There is a difference in scores for female and male instructors on items that assess the personal interaction component scores of the SET.

Experience Experience is a demographic factor that could affect teaching performance and thus the SET. Research suggests that experience does improve an instructor's SET. Ratz (1975) reported a 6% increase in SET score if a course has been taught previously by that instructor. Ragan and Walia (2010) found that instructors have low SETs for their first two years of teaching. Other studies (e.g., Morgan & Davies, 2006) found that after the initial semester, instructor SET scores significantly improved. It should be noted that data in their study is rather unique: instructors were typically foreign and teaching for the first time in an all-female Arab university. McPherson (2006) used a broader sample, but had a more limited result; he found that experience improved the SET, but only for lower-level courses. These findings conflict with those of Marsh (2007) who found that the SETs of individual faculty did not change over time.

In addition to studying experience, some studies examined the role of age. While these two variables may be highly correlated, they are different. While experience generally improved SET scores; older instructors (up to age 55) had lower SET scores (McPherson & Jewell, 2007; McPherson et al., 2009). Clayson (2009) evaluated the effects of age on a more granular level and found that as instructor age increases, scores for knowledge improve; those for rapport decrease.

We propose the following hypotheses for experience:

H_{5a} There is a positive correlation between experience and SET.

H_{5b} There is an increase in SET scores with increasing experience for assistant professors.

Academic rank As with experience, it is possible that instructors who better facilitate student learning are more likely to get promoted and therefore should have higher aggregate SET scores. Assistant professors may focus more on research in an attempt to earn tenure, while full professors may be better equipped to excel at both teaching and research (Feldman, 1987). A higher correlation between research productivity (as measured by publications) and student ratings of instructor effectiveness is seen after seven years of experience (a common time frame for tenure) for professional areas of study such as engineering (Centra, 1983).

Previous research on academic rank and SETs is inconclusive. Some studies of rank have shown no differences among faculty of differing rank (Aleamoni & Hexner, 1980; Troy et al., 1992; Zabaleta, 2007). Spooren (2010) found that associate professors and lecturers received statistically significantly lower SET scores than did full professors. Also supporting the positive correlation between rank and SET scores are findings that show tenure leads to higher SET scores (McPherson & Jewell, 2007; McPherson et al., 2009). Research has also compared the performance of graduate student faculty with permanent instructors. Blackhart et al. (2006) found that non-tenure-track faculty (lecturers and teaching assistants) received higher SET ratings, with teaching assistants performing the best. This result conflicts with other results that show performance of permanent faculty and teaching assistants is approximately the same (Schuckman, 1990). Another differentiation is between full-time and part-time faculty (i.e., adjunct professors excluding teaching assistants). Part-time faculty received higher SET scores (Goldberg & Callahan, 1991; McPherson & Jewell, 2007; McPherson et al., 2009); although in some cases, they also gave higher grades (Goldberg & Callahan; 1991). Peterson et al. (2008) found that full-time faculty gave higher grades than their part-time or non-permanent equivalents. According to Goldberg and Callahan (1991), there is a perception that higher grades lead to higher SET scores and that non-tenure-track faculty may give higher grades and, therefore, have higher SET scores.

These inconclusive results and perceptions led to the following hypotheses:

H_{6a} Tenured faculty (associate and full professors) have higher SET scores than their non-tenured counterparts (assistant professors and lecturers).

H_{6b} Tenure-track faculty (assistant professors) have lower SET scores than their non-tenure-track counterparts (lecturers and PhD students).

H_{6c} The average grade given by non-tenure-track faculty is higher than that given by tenure-track faculty.

H_{6d} There is a higher correlation between course grade and SET score for non-tenure-track faculty than for their tenure-track colleagues.

Student Perception of Assessment

Many factors influence students' perception of a course. These factors are commonly highly correlated (Peterson et al., 2008), and students do not always distinguish well between how the various factors influence their overall evaluation of teaching. Indeed, part of this study is to distinguish as well as possible how various factors may be influencing the overall SET of a course. Grades in particular have been studied so extensively and are of significant enough concern that they warrant a specific discussion here.

Actual or expected grade One of the most controversial and pressing concerns about SETs is the relationship between SET score and student grades (Boysen, 2008). Many investigators have analyzed this relationship and come to significantly different conclusions. Some have deemed the search for biases in SET relationships – such as that between grades and SET – a "witch hunt" (Marsh, 1984); others have termed the validity of the SETs as a measure of teacher effectiveness a "myth" (Johnson, 2003).

Three major reasons are usually offered for any correlation between student grades (either expected or actual) and SETs: grading leniency – higher grades being rewarded by higher SETs; validity – better teaching producing higher grades and higher SET; and student

characteristics – students' interest in a course leading to higher grades and higher SETs (Marsh, 1984). As mentioned previously, Cohen (1981) found a statistically significant correlation between objective measures of student achievement and SET scores.

Significant evidence of the positive correlation between course grades and SETs exists. This correlation has been shown for expected course grades (Aleamoni & Hexner, 1980), actual course grades (Blackhart et al., 2006), and both expected and actual course grades (Langbein, 2008). Ellis et al. (2003) reported that this relationship is more significant for the instructor than for the overall course. Adjunct faculty were shown to give better grades and receive better SET scores (Goldberg & Callahan, 1991). These correlations have been shown in disciplines as varied as the geosciences (Hippensteel & Martin, 2005), optometry (Trick, Lehmkuhle, Myers, Graham, & Davis, 1993), and education (DuCette & Kenney, 1982). Other cross-discipline studies have also shown these correlations (Kapel, 1974; Kidd & Latif, 2004; Isely & Singh, 2005). These correlations have been reported in the United States (McPherson & Jewell, 2007), New Zealand (Watkins, 1990), and the United Arab Emirates (Morgan & Davies, 2006).

Several studies acknowledge a relationship between grades and SET, but offer reasons relating either to the validity or student characteristic frameworks; for instance, that students have learned more (Cohen, 1989) or are more motivated (Howard & Maxwell, 1980), respectively. Howard and Maxwell (1982) also point out that the causality of the relationship is no more likely for leniency than validity. Marsh and Roche (1997) highlighted the small effect size of the correlations and noted the lack of causal evidence. Kherfi (2011) showed that students with higher overall cumulative grade averages (henceforth GPA) are more likely to participate in SETs and thus bias ratings upward. Spooren (2010) also showed that SET scores are correlated with GPA; better students may be learning more and providing higher SET scores. Marsh and Roche (2000) showed that when controlling for student characteristics, the remaining effect of grades on SETs is small. Patrick (2011) controlled for the amount of student learning and also found the strength of the course grade–SET relationship to be insignificant. Structural equation modeling also showed SET scores were based more on student stimulation and course enjoyment as opposed to the grade received (Remedios & Lieberman, 2008). Zabaleta (2007) showed a moderate correlation for lower grades leading to lower SET scores, but no relationship between higher grades and higher SET scores.

Some researchers have attempted to establish reciprocity or quid pro quo explicitly. Clayson, Frost, and Sheffet (2006) used a controlled situation to show that when instructors gave higher grades later in the semester, students gave higher SET scores. Students gave lower SET scores with lower grades even if they were told that other students were doing well in the course (Boysen, 2008). Anonymity also plays a role in the relationship between grades and SETs; anonymous students gave lower SETs than those that believed they might be identifiable (Blunt, 1991).

Other research has highlighted the relationship between perceived fairness and SET. Holmes (1972) highlighted the effect of the difference between expected and actual grades and SET. This difference suggests that students punish instructors who they do not believe are assessing them fairly. Bunker and Clayson (2008) explicitly highlighted the propensity of students to punish instructors if they feel they deserved a higher grade. The difference between the expected grade in a course and the GPA is significantly correlated with SET scores; a student who is receiving a higher grade than his or her GPA is likelier to give a higher SET score (Isely & Singh, 2005). Rodabaugh and Kravitz (1994) used scenarios presented to students to show that while course grades were correlated with SET scores,

instructor fairness had a stronger effect. Perceived grading difficulty (Pohlmann, 1975; Petchers & Chow, 1988) and strict grading (Nimmer & Stone, 1991) also resulted in lower SET scores. The following hypotheses are proposed to evaluate the relationship between SET and grades in light of the previous research:

H_{7a} There is a statistically significant positive correlation between SET scores and the course grade.

H_{7b} There will be a lower correlation between course grades and SET scores in small classes where students are likely to feel more identifiable.

H_{7c} Instructors receiving a low score for the fairness in grading item will have a higher correlation between grades and SET score (excluding that item) than those who received a higher score for fairness in grading.

H_{7d} Courses with a higher course grade (greater than 3.25/4.00) will have a higher correlation between grades and SET score than those courses with a lower average grade (less than 2.75/4.00).

Methods

The data for our analysis comes from seven consecutive semesters from fall 2007 to spring 2010 and spans the 11 major departments solely in the Dwight Look College of Engineering at Texas A&M University's College Station campus. The Department of Biological and Agricultural Engineering is omitted given its additional inclusion in the College of Agriculture and Life Sciences. The Dwight Look College offers undergraduate degrees in all 11 departments, and master's and doctoral degrees in all departments except the Department of Engineering Technology and Industrial Distribution. The students complete a teaching evaluation instrument towards the end of the semester. The instrument is either administered by paper in the classroom or online through a common system at the university. Data were collected for all courses (419 undergraduate and 397graduate) taught during that time. Some courses were taught every semester, while others were offered in alternate or intermittent semesters. They were also taught in multiple sections by the same or different faculty members depending on their class size and course load of the individual faculty members. In total, this analysis includes data for 3938 courses, and the overall enrollment in these courses was 137,431 students. Faculty of different rank teach these courses; in this dataset, 549 distinct faculty members (25 PhD students, 79 lecturers, 141 assistant professors, 121 associate professors, and 183 professors) have taught courses. Tables 2 and 3 (p. 300) summarize the data used in this study.

Course and Faculty Characteristics

The course type and class level were obtained from the course catalog and degree plans available on the department Web sites; the class size for each semester was obtained from the registrar's office. Faculty characteristics were obtained from CVs, department Web sites, and public databases. Experience is the number of years since each faculty member received his or her terminal degree; within the dataset, experience is coded to increase from fall 2007 to spring 2010 so that it gives a truer representation of the instructor's experience in the classroom over the study period.

SET Instrument

The SET instrument administered in the College of Engineering has the following eight questions:

Q1. *Class Preparation*: The class activities are well prepared and organized.

Q2. *Assignments*: The examinations, assignments, projects, etc. aid me in achieving the class objectives.

Q3. *Communications*: The instructor clearly explains material so that I can understand it.

Q4. *Responsiveness*: The instructor is open to my questions and effectively answers them.

Q5. *Academic Concern*: The instructor seems to care that I learn this material.

Q6. *Availability*: The instructor willingly makes time to help other students and me.

Q7. *Fairness in Grading*: The instructor is fair and consistent in evaluating my performance in the course.

Q8. *Environment*: The instructor maintains a good learning environment for me.

The students rate each question on a scale of 1 to 5:

1. Has serious deficiencies in the area which are detrimental to students
2. Does not perform well in this area
3. Good
4. Very good
5. Deserves an award in this area, excellent

The average score of each question per course section is used in this analysis. These are the same metrics used in evaluating faculty during their annual reviews as well as tenure and promotion at this institution.

Validity of the SET A valid SET must measure delivery, depth of knowledge, interpersonal skills, organization, and relevancy (Divoky & Rothermel, 1988). The questions in this SET measure some of these factors directly; for example, Q1 addresses preparation and organization, and Q3 addresses communication. Also included are availability and fairness in grading, Q6 and Q7, respectively. Brightman (2005), Centra (1982), and Peterson et al. (2008) claimed that when such factors are present in the SET, the perception that evaluation is a mere popularity contest is effectively refuted. To assess the validity of the SET, we conducted correlation, exploratory factor, and reliability analyses.

Table 1 presents the pairwise correlation matrix of the eight questions in the SET for the 3938 classes. Of note are the high correlations between each pair of items and the overall average: the lowest correlation is 0.707 and is between Q1 and Q6, which is still highly statistically significant.

To measure the dimensionality of the instrument, an exploratory factor analysis (EFA) on the SET instrument based on principal components was performed. The overall EFA

Table 1 Correlation of SET Instrument Questions

	Overall average	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
Overall average	–								
Q1	0.876	–							
Q2	0.899	0.799	–						
Q3	0.916	0.817	0.805	–					
Q4	0.930	0.767	0.780	0.852	–				
Q5	0.906	0.724	0.765	0.785	0.842	–			
Q6	0.888	0.721	0.744	0.725	0.839	0.845	–		
Q7	0.886	0.707	0.793	0.749	0.794	0.782	0.781	–	
Q8	0.943	0.803	0.826	0.864	0.863	0.844	0.812	0.823	–

Note. All correlations are significant at 0.01.

analysis included data from all 3938 courses, and it yielded a single factor with an eigenvalue of 6.571, and it accounted for 82.1% of the variance. The eigenvalue rule of Kaiser (1960) and scree test of Cattell (1966) were used to determine the number of components in the factor analysis. The eigenvalues of the first five factors were 6.57, 0.39, 0.27, 0.22, and 0.18. To reconfirm, an EFA based on course level was conducted. For each course level, only a single factor that explained at least 74% of the variance was obtained. The eigenvalues (variances) for freshman, sophomore, junior, senior, and graduate classes were 6.884 (86%), 6.392 (79%), 6.649 (83%), 6.450 (80%), and 5.926 (74%), respectively; all factor loadings in the analysis were above 0.8.

Reliability analysis was conducted to further evaluate the reliability of the instrument. Cronbach's α was obtained for each course level and globally for the entire dataset. They were 0.975 (freshman), 0.962 (sophomore), 0.970 (junior), 0.965 (senior), 0.949 (graduate), and 0.968 (overall). This strong reliability confirms the unidimensionality of the instrument. A similar result is seen in the Montclair State University's business school dataset (Peterson et al., 2008). To assess the construct validity of the instrument, Peterson et al.'s (2008) content validity method was used.

This SET instrument was developed by the faculty members of the College of Engineering and has been used as the primary quantifying metric for the last 15 years in evaluating faculty members' teaching performance during their tenure and promotion (T&P). Other qualitative metrics such as peer evaluation and educational development activities are included in the T&P package, but the score is the only standard quantitative metric available for measurement. The SET could be a one-time snapshot of what happened that day or week of the class. Despite this possible detriment, this SET is the only available quantitative metric at the study institution.

Grade Distribution

In almost all higher education institutions in the United States, the grades are assigned in terms of letters (A, B, C, D, or F); an A is equated to 4.0, B to 3.0, and so on. A student GPA of 4.0 is considered to be outstanding. Some institutions provide additional grade differentiation such as A+, A−, and so on. At the study institution only five distinct grades are used: 4.0 (A), 3.0 (B), 2.0 (C), 1.0 (D), 0.0 (F). Individual student grades or expected grades are not recorded along with their evaluation; this is done to maintain the anonymity

Table 2 Summary of Departmental and Faculty Data

	Faculty			% Courses taught by				
	n	FEM	YRS	PHD	LECT	ASST	ASSO	PROF
Aerospace	37	8%	19.5		19	14	22	46
Biomedical	20	20%	15.8		5	35	35	25
Chemical	38	18%	16.7	3	26	32	13	26
Computer science	52	17%	15.9		17	23	23	37
Civil (includes ocean)	88	15%	18.6	1	10	26	32	31
Electrical and computer	73	12%	17.7		4	32	23	41
General engineering[a]	70	20%	19.6	3	44	9	16	29
Eng. technology & industrial distribution	42	10%	13.2	7	24	36	14	19
Industrial engineering	49	18%	13.3	31	10	16	16	27
Mechanical	84	14%	18.5	5	8	31	19	37
Nuclear	26	15%	20.5		12	31	23	35
Petroleum	36	14%	22.6	3	22	8	25	42
Total for College of Engineering	549	14%	17.5	5	14	25	22	33

Note. FEM = percent female; YRS = experience; PHD = doctoral student; LECT = lecturers; ASST = assistant professor; ASSO = associate professor; PROF = full professor.
[a]General engineering courses within the College of Engineering serve freshmen engineering, graphics, and ethics courses. Mostly faculty members from other engineering departments teach in this department. This department does not offer any degrees of its own.

Table 3 Summary of Departmental and SET and Course Grade Data

	SET		CG		CS		SET-CG
	M	*SD*	*M*	*SD*	*M*	*SD*	*r*
Aerospace	4.02	0.51	3.31	0.56	29.4	18.2	0.33*
Biomedical	4.18	0.44	3.62	0.37	27.3	26.8	−0.034
Chemical	4.26	0.44	3.30	0.45	34.6	24.2	0.372*
Computer science	4.15	0.48	3.34	0.53	32.9	30.3	0.463*
Civil (includes ocean)	4.29	0.44	3.27	0.50	35.0	24.8	0.431*
Electrical and computer	4.29	0.43	3.33	0.50	24.8	16.8	0.474*
General engineering[a]	3.82	0.52	3.11	0.40	54.6	36.6	0.314*
Eng. technology & industrial distribution	4.26	0.46	3.08	0.50	35.2	20.8	0.376*
Industrial engineering	4.35	0.37	3.38	0.36	31.0	17.4	0.388*
Mechanical	4.12	0.47	3.18	0.53	39.9	30.3	0.345*
Nuclear	4.35	0.38	3.52	0.45	21.5	14.9	0.352*
Petroleum	4.07	0.45	3.32	0.43	47.0	38.5	0.431*
Total for College of Engineering	4.19	0.48	3.38	0.50	34.8	27.1	0.374*

Note. SET = student evaluation of teaching score; CG = course grade; CS = class size; SET-CG = correlation of SET and CG.
[a]General engineering courses within the College of Engineering serve freshmen engineering, graphics, and ethics courses. Mostly faculty members from other engineering departments teach in this department. This department does not offer any degrees of its own.
*$p < 0.001$.

of the evaluation process. The overall grade distribution of each course was obtained from the registrar's office at the university. The dataset includes the grades assigned by the faculty, along with the course name, number, section, and semester. For this analysis, the grade distribution dataset is merged with the respective classroom evaluation, class characteristic, and instructor characteristic details to obtain the final data for analysis. Descriptive demographic and SET data for the study are given in Tables 2 and 3, respectively.

Results

Course Characteristics

The effects of the course characteristics on the SET were examined using statistical analyses. An alpha level of 0.05 was used for all statistical tests. Data were examined to assess the effects of class size, course level, and course type.

Class size To examine the relationship between course size and SET, three hypotheses were proposed. The first related to the correlation between course size and SET (unless otherwise noted, SET refers to the average of the eight question scores). The statistically significant negative correlation between aggregate SET and class size ($N = 3938$, $r = -0.291$, $p < 0.001$) suggests that faculty who teach smaller classes receive higher SET scores and that there is insufficient evidence to reject hypothesis H_{1a}. Table 4 shows aggregate SET scores for various class sizes. Aggregate SET seems to decrease monotonically with increasing class size. Therefore, there is no evidence of a parabolic relationship between SET and course size, and hypothesis H_{1b} is rejected. Table 5 shows the correlations between course size and SET as categorized by course level. The greatest negative correlation occurs in the junior-level courses. These data do not provide conclusive support for hypothesis H_{1c}. Fisher's z-test compares the correlations between class size and SET for freshman and sophomore courses ($r = -0.139$, $p < 0.001$) and junior and senior courses ($r = -0.175$, $p < 0.001$); the difference between the correlations was not significant ($z = 0.85$, $p = 0.400$).

Course level The next three course characteristic hypotheses tested examined the effects of course level on SET. The data show statistically significant support for hypothesis H_{2a} ($N = 3938$, $r = 0.341$, $p < 0.001$); there is a positive correlation between course level and SET. Hypothesis H_{2b} examined the role of experience on the SET at various course levels. The data shown in Table 6 partially support this hypothesis. Instructors with more years of experience had a higher aggregate SET for freshman-level courses. All other correlations between experience and the SET were negative; more-experienced faculty received lower SET scores. For sophomore- and graduate-level courses, these correlations were not significant. Fisher's z examined the correlations between the freshman-level courses and all other undergraduate courses. The correlation between experience and SET for the freshman courses was significantly higher than for other undergraduate courses. The final hypothesis tested related to course level examined the role of gender on aggregate SET at various course levels. A summary of previous work had presented conflicting data regarding this interaction (Feldman, 1993). The results in Table 7 indicate that female faculty receive significantly lower (2%) SET scores in lower-level courses (freshman and sophomore) than their male counterparts. The Welch t-test is used here (and in other similar instances), given the unequal sample sizes. Lower SET scores for female faculty are not the case in upper-level or graduate courses where there is no such difference. These results should be further emphasized because, while female faculty teach 17% of the courses in the dataset, they teach almost 25% of the lower-level courses.

Table 4 SET Scores by Course Size Range

Range	*n*	*M*	*SD*
1–20	1551	4.32	0.46
21–40	1114	4.18	0.44
41–60	669	4.09	0.47
61–80	335	4.01	0.46
81–100	148	3.98	0.48
101–150	106	3.82	0.42
151–200	15	3.44	0.39

Table 5 Correlation of SET Scores and Course Size by Course Level

Class level	*n*	*r*	*p*
Freshman (100 level)	305	−0.170	0.003
Sophomore (200 level)	384	−0.085	0.094
Junior (300 level)	906	−0.237	< 0.001
Senior (400 level)	1170	−0.073	0.012
Graduate (600 level)	1173	−0.141	< 0.001

Table 6 Correlation of SET Scores and Instructor Experience by Course Level

Course level	*n*	*r*	*p*	*z*[a]	*p*[b]
Non-freshman undergrad	2460	−0.081	< 0.001	2.99	0.003
Freshman (100 level)	305	0.101	0.078		
Sophomore (200 level)	384	−0.054	0.288		
Junior (300 level)	906	−0.112	0.001		
Senior (400 level)	1170	−0.087	0.003		
Graduate (600 level)	1173	−0.024	0.416		

[a]Fisher's *z* for comparison between non-freshman and freshman correlations.
[b]*p* for Fisher's *z*.

Table 7 Comparison of Male and Female Instructor SET Scores by Course Level

Class level	Test	Group	*n*	*M*	*t*	*p*
Upper	Male SET >Female SET	Male	2876	4.229	−0.21	0.582
		Female	373	4.233		
Lower	Male SET >Female SET	Male	518	3.998	1.66	0.049
		Female	171	3.932		

Note. Two-tailed *t*-test compares upper- and lower-level course SET scores by gender.

Course type The final course characteristic examined related to whether or not the course was required for the student's major. Two hypotheses were tested. The first examined the effect of elective vs. core or major courses. The data in Table 8 provide statistically significant support for hypothesis H_{3a}. The second hypothesis, H_{3b}, proposed that electives

would have a lower correlation between SET and course grades than core courses. The data show the opposite. Electives ($n = 2187$, $r = 0.400$, $p < 0.001$) had a higher correlation than core courses ($n = 1751$, $r = 0.298$, $p < 0.001$). The differences between the two correlations were tested using Fisher's z. There is a significantly stronger correlation between course grade and SET for electives ($z = -3.62$, $p < 0.001$). This is an unexpected result that contradicts some previous research (DuCette & Kenney, 1982).

Instructor Demographics

To analyze the relationship between instructor demographics and SET, the effects of gender, experience, and academic rank were examined. The eight hypotheses stated above were examined using statistical tests of the dataset.

Gender Hypothesis H_{4a} proposed a difference between male and female instructor SETs. The data in Table 9 support the hypothesis that male instructors receive higher SET scores than do their female counterparts. The second hypothesis examined whether female faculty were being held to a higher standard based on interpersonal aspects of teaching. Hypothesis H_{4b} examined the two SET questions most aligned with those aspects: academic concern (Q5) and availability (Q6). The data in Table 10 does not support this hypothesis. There is no statistically significant difference between male and female instructors for SET questions Q5 and Q6.

Experience Experience should have an effect on SET. The premise underlying the two hypotheses related to experience is that as instructors gain more experience, their teaching abilities should improve. The first hypothesis tested this assumption directly. However, the data do not support the claim that experience improves teaching ($N = 3938$, $r = -0.05$, $p = 0.001$). In fact, the correlation, while small, is negative and significant. To assess if there was an increase in SET with experience for assistant professors, the assistant professor subgroup was assessed separately. Again, there was a significant negative correlation between experience and SET score ($n = 923$, $r = -0.18$, $p < 0.001$); in the case of assistant professors, it was over three times greater than that of the entire teaching population.

Academic rank As with experience, it is reasonable to assume that faculty who have achieved more senior ranks are likely to be better teachers and thus receive higher SET scores. In addition to the rank of tenure-track faculty, the effect of non-tenure-track faculty (adjuncts, lecturers, and PhD students) status on SET scores was also assessed. Four hypotheses were tested. The first, H_{6a}, compared tenured faculty (associate and full professors) to their non-tenured counterparts (assistant professors, lecturers, and PhD students). The data shown for the first test in Table 11 do not support this hypothesis; tenured faculty do not receive significantly higher SET scores than their non-tenured counterparts. The second comparison for faculty rank compared tenure-track (assistant professors) to non-tenured and non-tenure-track faculty. The data shown for the second test in Table 11 shows that hypothesis H_{6b} is not supported. In fact, the opposite is supported; tenure-track assistant professors receive statistically significantly higher SET scores than non-tenure-track faculty. The next hypothesis examined the mean grades given by faculty of differing tenure-track status. The assumption was that non-tenure-track faculty would give higher grades and receive higher SET scores. Hypothesis H_{6c} was not supported. Again, the opposite was supported; non-tenure-track faculty assigned significantly lower grades than did their tenure-track and tenured counterparts. These data are shown in Table 12. The final hypothesis related to faculty rank assessed the correlations between grades and SET scores for tenure-track and non-tenure-track faculty. The correlations for tenure-track and non-tenure-track faculty were not

Table 8 Comparison of Elective and Core Courses

Test	Group	*n*	*M*	*t*	*p*
Elective > Core	Elective	2187	4.245	8.72	< 0.001
	Core	1751	4.112		

Note. Two-tailed *t*-test compares elective and core course types.

Table 9 Comparison of Male and Female Instructors

Test	Group	*n*	*M*	*t*	*p*
Male SET >Female SET	Male	3394	4.19	2.59	0.010
	Female	544	4.13		

Note. Two-tailed *t*-test compares SET scores by faculty gender.

Table 10 Comparison of Male and Female Instructors on Interpersonal Aspects of SET

Test	Group	*n*	*M*	*t*	*p*
Male SET >Female SET Q5	Male	3394	4.30	0.08	0.937
	Female	544	4.30		
Male SET >Female SET Q6	Male	3394	4.24	0.44	0.659
	Female	544	4.23		

Note. Two-tailed *t*-test compares SET questions Q5 and Q6 by faculty gender.

Table 11 Effect of Faculty Rank on SET Scores

Test	Group	*n*	*M*	*t*	*p*
Tenured > Non-tenured	Tenured	2344	4.191	0.86	0.195
	Non-tenured	1594	4.178		
Assistant professor > Non-tenure track	Assistant professors	923	4.237	5.86	<0.001
	Non-tenure track	671	4.096		

Note. Two-tailed *t*-test compares SET scores of tenured vs. non-tenured faculty and assistant professors vs. non-tenured faculty.

Table 12 Effect of Faculty Rank on Course Grade

Test	Group	*n*	*M*	*t*	*p*
Tenure track > Non-tenure track	Non-tenure track	671	3.18	5.89	<0.001
	Tenure track	3267	3.30		

Note: Two-tailed *t*-test compares course grades for non-tenure-track and tenure-track faculty.

Table 13 Correlation of SET Scores and Course Grades for Tenure-track and Non-tenure-track Faculty

Group	*n*	*r*	*p*	z^a	p^b
Non-tenure-track	671	0.370	<0.001	0.03	0.976
Tenure-track	3267	0.369	<0.001		
Assistant professor	923	0.410	<0.001		

[a]Fisher's *z* for comparison between non-tenure-track and tenure-track correlations.
[b]*p* for Fisher's *z*.

Table 14 Correlation of Aggregate SET and Course Grade for Varying Class Sizes

Class size	*n*	*r*	*p*	z^a	p^b
Small (<30)	2099	0.279	<0.001	−2.47	0.014
Medium (≥30 <60)	1213	0.359	<0.001		
Large (≥60 ≤150)	611	0.271	<0.001		
Very large (>150)	15	0.408	0.131		

[a]Fisher's *z* for comparison between small and medium size class correlations.
[b]*p* for Fisher's *z*.

Table 15 Correlation of Aggregate SET and Course Grade for Differing Level of Perceived Fairness in Grading Rating

Fairness in grading score	*N*	*r*	*p*	z^a	p^b
<4.0	1155	0.103	< 0.001	−4.86	< 0.001
≥4.0	2783	0.267	< 0.001		

[a]Fisher's *z* for comparison between lower (<4.0) and higher (≥ 4.0) fairness in grading scores.
[b]*p* for Fisher's *z*.

Table 16 Correlation of SET and Course Grade for Varying Course Grades

Course grade	*n*	*r*	*p*	z^a	p^b
≤2.75	675	0.188	<0.001	0.05	0.960
≥3.25	2149	0.186	<0.001		
2.75–3.25	1114	0.068	0.022		

[a]Fisher's *z* for comparison between lower (≤ 2.75) and higher (≥ 3.25) course grades.
[b]*p* for Fisher's *z*.

significantly different (the Fisher *z* shown in Table 13 compares the correlations between these two subgroups). The correlation between grades and SET for the tenure-track assistant professor subgroup was 11% higher than for non-tenure-tack instructors and PhD students; this difference is not statistically significant ($z = -0.93$, $p = 0.352$). There was no difference between the entire tenure-track faculty subgroup and the non-tenure-track faculty subgroup.

This is evidence for the rejection of hypothesis H_{6d}. This result again is in conflict with previous studies cited in the literature.

Student Perception of Assessment

Actual or expected grade As stated above, one of the most controversial and most studied aspects of SETs is the relationship between SETs and course grades. While some of the previous results have highlighted certain aspects of the relationship, and Table 3 shows the correlation between SET scores and course grade by department, specific aspects of the course grade and SET relationship are examined through the testing of three hypotheses. The first hypothesis, H_{7a}, tested the overall relationship between aggregate SET and course grades ($N = 3938$, $r = 0.374$, $p < 0.001$). These data support the hypothesis that there is a positive relationship between aggregate SET scores and course grades. The second hypothesis examined the role of anonymity on the course grade–SET relationship. It was assumed that if students were anonymous they would reciprocate lower course grades with lower SET scores. There was partial support for hypothesis H_{7b}; as shown in Table 14, there is a significantly lower correlation in small classes than medium-size classes. However, the correlation for course grades and SET scores in small courses is not significantly different from large ($z = 0.19$, $p = 0.849$) or very large courses ($z = -0.51$, $p = 0.610$).

The third hypothesis, H_{7c}, assessed the relationship between perceived fairness of grading and the course grade–SET relationship. It was assumed that instructors who received low grading fairness scores (Q7) would be more likely to be rewarded with higher SET scores by students who received higher grades. This was not the case. The results shown in Table 15 indicate that instructors who received higher SET scores for grading fairness had significantly higher correlations between SET scores and course grades.

Finally, the effect of average grade on the correlation between SET and course grades was examined. Hypothesis H_{7d} assumed that instructors who graded more leniently (with course grades $\geq 3.25/4.00$) would have a higher correlation between course grades than less-lenient instructors. There was no evidence to support this hypothesis because the correlations for both groups were similar; Fisher's z in Table 16 compares these two subgroups. However, there was a significantly lower correlation between course grades and SET scores for courses where the average grade was between 2.75 and 3.25 and those below 2.75 ($z = 2.50$, $p = 0.012$). There was also a significantly lower correlation for courses with average grades between 2.75 and 3.25 and those above 3.25 ($z = 3.25$, $p = 0.001$).

Summary of Hypothesis Tests and Sources of Variance

A summary of the various hypotheses and the statistics used to test them is shown in Tables 17–19. One-way analysis of variance (ANOVA) was used to examine the relative contributions of the course and instructor characteristics to variance of SET scores (see Table 20). As expected, course grade, class size, and average years of experience are not normally distributed. To eliminate the problem, we apply the appropriate Yeo and Johnson (2000) transformations on the data prior to performing the ANOVA. All course grade, course, and instructor variables (with the exception of gender) are significant. An examination of the marginal sum of squares, *SS* (Type III), shows that course grade contributes most to the variation among SET scores; higher grades are positively correlated with higher SET scores. Course grades are followed by the class level and class size; higher SET scores are seen in senior and smaller-size classes. The overall amount of variation in SET scores for the main effects shown in Table 20 is 21.6% (adjusted R^2). When taking into account two-way

Table 17 Summary of Course Characteristics Hypotheses and Results

Course characteristics hypotheses	Results	Sample size (number of classes)	Tests
Class size & SET			
H_{1a} Negative linear correlation	Supported***	3938	Pearson correlation
H_{1b} Parabolic relationship	Not supported	3938	Graphical (visual) analysis
H_{1c} Upper-level (lower correlation) vs. lower-level (higher correlation) courses	Not supported	*FR* (305); *SO* (384); *JR* (906); *SR* (1170); *GR* (1173)	Fisher's z
Course level & SET			
H_{2a} Positive linear correlation	Supported***	3938	Pearson correlation
H_{2b} Relationship based on instructor experience, higher in lower-level courses than upper-level courses	Partially supported**	*FR* (305); *SO* (384); *JR* (906); *SR* (1170); *GR* (1173)	Fisher's z
H_{2a} Relationship based on instructor gender	Supported in lower-level courses only*	*MHL* (2876); *FHL* (373); *MLL* (518); *FLL* (171)	Welch t
Course type & SET			
H_{3a} Elective courses have higher SET than core courses	Supported***	*EL* (2187); *CR* (1751)	Welch t
H_{3b} Correlation higher in core courses than in elective courses	Not supported (opposite is supported)***	*EL* (2187); *CR* (1751)	Fisher's z

Note. *FR* = freshman; *SO* = sophomore; *JR* = junior; *SR* = senior; *GR* = graduate; *MHL* = male instructor, higher-level course; *FHL* = female instructor, higher-level course; *MLL* = male instructor lower-level course; *FLL* = female instructor lower-level course; *EL* = elective course; *CR* = core course.
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

interactions as well as main effects, instructor rank as well as its interactions with instructor experience, average course grade, and course level accounted for the four largest marginal sum of squares values. The fifth largest was the interaction between course type and course level. These results are shown in Table 21 and further confirm the hypotheses and results above. The overall amount of variance accounted for using both main effects and two-way interactions is 26.2% (adjusted R^2).

Close examination of the two-way ANOVA (Table 21) shows that each of the main effects was part of at least two separate statistically significant two-way interactions. (Course Type, 2; Class Size, 3; Instructor Experience, 3; Instructor Gender, 3; Course Grade, 4; Instructor Rank, 4; and Class Level, 5). Five of these interactions were specifically addressed in the original hypotheses. Course size and course level exhibit a complex relationship as shown in Table 5; there seems to be no trend in the relationship, but the highest correlation between SET and size is seen in junior classes. Class level and instructor gender are shown to interact negatively at lower course levels for female instructors, while there is little difference in upper-level courses, as shown in Table 7. Instructor experience has a positive correlation with SET scores for freshman courses and is consistently (though not always significantly) negatively correlated at upper-level courses, as seen in Table 6. GPA leads to a

Table 18 Summary of Instructor Characteristics Hypotheses and Results

Instructor characteristics hypotheses	Results	Sample size (number of classes)	Tests
Gender & SET			
H_{4a} Male instructors receive higher SET than female instructors	Supported**	*M* (3394); *F* (544)	Welch *t*
H_{4b} Difference with respect to personal interaction items in SET instrument	Not supported	*M* (3394); *F* (544)	Welch *t*
Experience & SET			
H_{5a} Positive linear correlation	Not supported (small negative correlation)***	3938	Pearson correlation
H_{5b} Assistant professor subgroup only	Not supported (sig. negative correlation)***	923	Pearson correlation
Academic Rank & SET			
H_{6a} Tenured faculty have higher SET than non-tenured counterparts	Not supported	*TF* (2344); *NTF* (1594)	Welch *t*
H_{6b} Assistant professors have lower SET than non-tenure-track counterparts	Not supported (Opposite is supported)***	*ASST* (923); *NTT* (671)	Welch *t*
H_{6c} Non-permanent instructor give higher CG	Not supported (Opposite is supported)***	*TT* (3267); *NTT* (671)	Welch *t*
H_{6d} Correlation between SET and CG is higher for non-tenure-track faculty	Not supported	*TT* (3267); *NTT* (671)	Fisher's *z*

Note. *M* = male instructor; *F* = female instructor; *TF* = tenured faculty member; *NTF* = non-tenured faculty member; *ASST* = assistant professor; *NTT* = non-tenure-track faculty; *TT* = tenure-track faculty; *CG* = course grade.
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

higher SET score in medium-size classes (30–60 students) than it does in smaller and larger courses (Table 14). Experience is negatively correlated with SET scores, specifically for assistant professors, as seen in the results. This negative correlation might indicate that over time assistant professors are realigning their priorities away from teaching as they move towards tenure.

Seven other significant interactions were not specifically tested as part of the original hypotheses, but the subsequent preliminary analyses were done for these other significant interactions. (Detailed data are not shown to save space and preserve readability.) Increasing experience was positively correlated with SET scores for non-tenure-track faculty, while negatively correlated for assistant professors and even more negatively correlated for tenured associate professors, while there was no correlation between experience and SET score for full professors. There is a larger negative correlation between SET scores and course size in

Table 19 Summary of Assessment and SET Hypotheses and Results

Assessment & SET hypotheses	Results	Sample size (number of classes)	Tests
H_{7a} Higher aggregate SET with higher grades	Supported**	3938	Pearson correlation
H_{7b} Lower correlation in smaller classes	Partially supported*	*SM* (2099); *MD* (1213); *LG* (611); *VL* (15)	Fisher's *z*
H_{7c} Lower fairness in grading (Q7) equates to higher correlation between grades and SET	Not supported (opposite is supported)**	*HFG* (2783); *LFG* (1155)	Fisher's *z*
H_{7d} Higher CG classes have higher correlation than lower CG classes	Not supported	$CG < 2.75$ (675); $CG > 3.25$ (2149)	Fisher's *z*

Note. *SM* = small courses; *MD* = medium courses; *LG* = large courses; *VL* = very large courses; *HFG* = high fairness in grading rating; *LFG* = low fairness in grading rating; *CG* = course grade.
*$p < 0.05$; **$p < 0.001$.

Table 20 One-Way ANOVA Results

Source	*F*	*P*	*SS*	*df*	*MS*
Corrected Model	83.779	< 0.001	183.213	13	14.093
Course Grade[a]	239.380	< 0.001	40.268	1	40.268
Class Size[b]	31.661	< 0.001	5.326	1	5.326
Instructor Experience[c]	10.205	0.001	1.717	1	1.717
Course Level	45.789	< 0.001	30.810	4	7.703
Instructor Gender	3.202	0.074	0.539	1	0.539
Instructor Rank	2.530	0.039	1.703	4	0.426
Course Type	3.898	0.048	0.656	1	0.656
Intercept	4182.626	< 0.001	703.598	1	703.598
Error			655.046	3894	0.168
Total			69526.324	3908	
Corrected total			838.259	3907	
R-square	0.219				
Adj. R-square	0.216				

[a]Transformed Course Grade: $((\text{Course Grade} + 1)^2 - 1)/2$.
[b]Transformed Class Size: ln (class size+1).
[c]Transformed years of experience: $2 * (\sqrt{\text{Instructor Experience}+1} - 1)$.

elective than in core courses. For course grades and course level there is a higher positive correlation between course grades and SET for freshman, sophomore, and junior courses than for senior and graduate courses (although it still positive for senior and graduate courses). Course grades have a larger positive correlation with SET scores for female instructors than for male instructors. For the interaction of course size with experience, experience was negatively correlated with SET for small, medium, and large classes; however, very large classes had no statistically significant correlation between experience and SET. The significant interactions for course type with class level and class level with instructor rank were comprised of two sets of categorical variables; the effective treatment means for the various factor levels were compared. The interaction of course type with class level showed that lower-level

Table 21 Combined One-Way and Two-Way ANOVA Results

Source	*F*	*p*	*SS*	*df*	*MS*
Corrected Model	19.266	< 0.001	231.790	76	3.050
Course Grade[a]	3.135	0.077	0.496	1	0.496
Class Size[b]	2.155	0.142	0.341	1	0.341
Instructor Experience[c]	0.119	0.730	0.019	1	0.019
Class Level	2.138	0.074	1.354	4	0.338
Instructor's Gender	0.190	0.663	0.030	1	0.030
Instructor's Rank	8.579	< 0.001	4.074	3	1.358
Course Type	0.380	0.538	0.060	1	0.060
Course Level * Class Size[b]	1.825	0.121	1.156	4	0.289
Course Level * Course Type	5.922	< 0.001	3.750	4	0.937
Course Level * Course Grade[a]	3.220	0.012	2.039	4	0.510
Course Level * Instructor Rank	2.400	0.002	5.319	14	0.380
Course Level * Instructor Gender	2.453	0.044	1.553	4	0.388
Course Level * Instructor Experience[c]	2.829	0.023	1.792	4	0.448
Course Type * Class Size[b]	5.037	0.025	0.797	1	0.797
Course Grade[a] * Class Size[b]	4.431	0.035	0.701	1	0.701
Instructor Rank * Class Size[b]	0.739	0.565	0.468	4	0.117
Instructor Gender * Class Size[b]	1.439	0.230	0.228	1	0.228
Instructor Experience[c]* Class Size[b]	2.747	0.098	0.435	1	0.435
Course Type * Course Grade[a]	0.075	0.784	0.012	1	0.012
Course Type * Instructor Rank	0.402	0.752	0.191	3	0.064
Course Type * Instructor Gender	1.275	0.259	0.202	1	0.202
Course Type * Instructor Experience[c]	0.002	0.967	0.000	1	0.000
Instructor Rank * Course Grade[a]	9.064	< 0.001	5.740	4	1.435
Instructor Gender * Course Grade[a]	3.541	0.060	0.560	1	0.560
Course Grade[a] * Instructor Experience[c]	0.108	0.742	0.017	1	0.017
Instructor Rank * Instructor Gender	5.520	< 0.001	3.496	4	0.874
Instructor Rank * Experience[c]	13.908	< 0.001	6.605	3	2.202
Instructor Gender * Experience[c]	0.958	0.328	0.152	1	0.152
Intercept	282.727	< 0.001	44.757	1	44.757
Error			606.468	3831	0.158
Total			69526.324	3908	
Corrected total			838.259	3907	
R-Square	0.277				
Adj. R-Square	0.262				

[a]Transformed Course Grade: $((\text{Course Grade} + 1)^2 - 1)/2$
[b]Transformed class size: ln (class size+1)
[c]Transformed years of experience: $2 * (\sqrt{\text{Instructor Experience} + 1} - 1)$

core courses had higher SET scores than did electives, while for senior and graduate courses there was an opposite effect, where electives had higher SET scores than core courses (junior courses were the same for core and elective). For the interaction of class level with instructor rank, assistant professors had lower SET scores in freshman courses, with increasing SETs with increasing course level (with the highest SET for graduate courses). Associate professors were slightly below average in freshmen courses, and near their group average at other course levels. PhD students were well below average for freshmen courses with increasing SET scores through senior-level courses; instructors were near average for all lower-level courses, but above their cohort average in senior-level courses. Full professors were above

average in freshman courses and slightly lower at all other levels; this may imply that having experienced full professors in freshman courses might provide a better student experience.

Discussion

The first set of effects on SETs examined in this study was course characteristics. Class size was found to have a significant negative correlation with SET scores. Faculty assigned to teach large courses are likely to receive lower aggregate SET scores. At the extremes (as seen in Table 4), instructors assigned to large courses (>100 students) had aggregate average SET scores ($M = 3.81$, $SD = 0.44$) that were more than one-half point lower than those of their peers assigned to small courses (< 20 students; $M = 4.32$, $SD = 0.46$). Overall, the results reported in this article are in broad agreement with the majority of the cited literature (Ellis et al., 2003; Burns & Ludlow, 2005; Westerlund, 2008; Chapman & Ludlow, 2010; Ragan & Walia, 2010).

Another significant variable that affected aggregate SET scores was course level. Upper-level courses received significantly higher SET scores. Experience also played a significant role at various course levels. While experience and aggregate SET scores were positively correlated at the freshman level, at other levels the correlations were either negative, insignificant, or both. Junior or inexperienced faculty assigned to teach freshman courses could be at a disadvantage and receive significantly lower SET scores. There is some support in the cited literature for this negative effect (Hippensteel & Martin, 2005). A particular effect of course level examined in this article was the comparison of male and female aggregate SET scores at lower- and upper-level courses. Female faculty had significantly lower SET scores in lower-level courses. This result agrees with a study cited by Feldman (1993).

The effect of required or core elective courses was also examined. The SET scores for elective courses were significantly higher than those of core courses. This finding agrees with several previous cited studies (Lovell & Haner, 1955; Kapel, 1974; Aleamoni & Hexner, 1980; Isely & Singh, 2005). While the hypothesis related to core and elective grade correlations was not supported, the differences between aggregate SET scores for the two types of course should be taken into account during the interpretation of SETs. Instructors assigned to core or required courses are at a disadvantage, and steps should be taken to ensure that faculty are given an opportunity to teach both types of course over any given evaluation period.

Another factor examined was instructor gender. Overall, male faculty received higher aggregate SET scores than did their female counterparts. However, as seen in Table 7 this was most significant for lower-level courses and not significant for upper-level courses. Again, the interpretation of female faculty members' SET scores for lower-level courses should be taken into account. In contrast to the findings of Clayson (2009), students in this study did not seem to hold female instructors to higher standards with respect to interpersonal-related SET questions.

The next demographic factor assessed was instructor experience. For the entire faculty population, a statistically significant negative correlation was found between experience and aggregate SET. This overall result agrees with some cited work that shows that older instructors receive lower SET scores (McPherson & Jewell, 2007; McPherson et al., 2009). One possible explanation for this decline could be that as faculty gain more experience teaching a particular course or teaching in general, they may take less time for preparation and thus receive lower SET scores.

The results for the assistant professor subgroup were unexpected. There was a greater significant negative correlation for this subgroup than the faculty as a whole (−0.18 vs. −0.05,

respectively). This could result due to the lower teaching load typically given to new faculty members. As these faculty members have to juggle more courses, their SET might decrease. Another possible explanation is that as faculty near their tenure and promotion year, their focus could shift more to research as opposed to teaching. This particular finding is a promising area for future research.

The final demographic variable analyzed was academic rank. While it was expected that tenured faculty would have higher aggregate SET scores than their non-tenured counterparts, this was not the case. This finding is in disagreement with previous cited work that showed tenure positively correlated with SET scores (McPherson & Jewell, 2007; McPherson et al., 2009). Assistant professors had higher aggregate SET scores than lecturers and PhD students. The finding that tenure-track faculty, not their non-tenure-track counterparts, give higher average grades (see Table 12) agrees with Peterson et al. (2008) and disagrees with the older work of Goldberg and Callahan (1991). There was no significant difference between the course grade and SET correlations for tenure-track and non-tenure-track faculty; however, the assistant professor subgroup's correlation was approximately 11% higher (though the difference is not statistically significantly different).

The final aspect of SET analyzed was the controversial relationship between course grades and SET. The results showed statistically significant positive correlations for all departments save one and a statistically significant positive correlation overall (see Table 3). This correlation was slightly higher than those cited in previous research (0.37 for this study vs. 0.10 to 0.30 for previous research [Cohen, 1989; Cashin, 1990; Braskamp & Ory, 1994; Marsh & Roche, 1997]). Other results in this study were less conclusive. There was no difference in correlation based on class size. Also, contrary to the proposed hypothesis, instructors receiving poor fairness in grading scores had higher correlations between SET and average course grades. Also unexpected was the bifurcation of correlations at the high (>3.25) and low (<2.75) ends for average course grade. These instructors had higher SET average course grade correlations than instructors whose average course grades were between these two levels. At a minimum, those assessing teaching effectiveness should evaluate the results for individual courses in light of average course grades across different instructors and take into account other observable learning outcomes.

The absolute effect of the variables discussed above should be assessed within the context of the evaluation instrument under consideration to determine its effect size (Marsh, 1984). While it is possible for instructors to score between 0 and 5 on their overall SET score, the mean for the entire College of Engineering is 4.19, as opposed to 3.0 – an inflation common among SETs (Cashin, 1990). Using the standard deviations, we can measure the effect sizes of differences between subgroups with Cohen's (1992) d. The effect sizes for both gender in lower-level ($d = 0.125$) and core vs. elective ($d = 0.213$) courses would be considered "small" (Cohen, 1992). However, the effect sizes of the correlations between SET and class size ($r = -0.291$), course level ($r = 0.341$), and course grade ($r = 0.374$) would all be considered "medium." While these effects are not individually large, they should be taken into account when evaluating SET as part of faculty teaching performance.

Conclusions

Given the importance of SETs in promotion, tenure, and retention decisions, it is important that SET scores be evaluated in light of factors that may not be directly related to teaching effectiveness. Considering these factors becomes even more important as interest in SETs

grows outside of academe. This study has shown that several course, instructor, and related factors have statistically significant effects on SET scores in a large public college of engineering. While the sample size of the presented data is relatively large, these results represent only one campus. Other colleges and programs should carry out their own analyses prior to generalizing the above results.

There was a significant negative correlation between course size and aggregate SET score. There was also a significant positive correlation between course level and aggregate SET. Faculty assigned to lower-level courses that tend to have large sections could be disadvantaged because they are likely to receive lower SETs, other things being equal. Administrators should take course characteristics into account when evaluating faculty performance in these types of course. Female faculty who teach lower-level courses are also likely to receive slightly lower SET scores, while at the upper level, there is no statistically significant difference. In this study, female faculty were overrepresented in their instruction of lower-level courses (based on total number of courses taught). Even though the effect size was small, those with administrative responsibility for course assignments should be cognizant of these results and assign faculty accordingly, or at a minimum take this into account when evaluating SET results.

Another significant difference in SET scores arises between core and elective courses. Instructors teaching elective courses received an average SET score 0.12 greater than those teaching core courses. Again, even in light of the small effect size, these differences should be noted. Faculty who consistently teach elective courses could have inflated SETs compared with those continuously assigned to core courses.

Finally, the 0.37 correlation between SET score and course grade is higher than the often reported range of 0.1 to 0.3 for the correlation between expected grade and SET score (Cohen, 1989; Cashin, 1990; Marsh & Roche, 1997); these higher SET scores should be further investigated and validated as representative of instructor performance. Faculty members who assign average course grades that are 0.5 point higher than those given by their colleagues could theoretically increase their aggregate SET scores by almost 0.2 point. Administrators and senior faculty should take course and instructor characteristics and course grades into account when evaluating faculty teaching effectiveness using SETs. In some cases it may be necessary to control SET results for biases associated with some characteristics, but decision makers should be careful not to give too much influence to these biases without firm statistical evidence (Cashin, 1990). This article has attempted to provide some of that statistical evidence in an engineering context.

Acknowledgement

The authors would like to thank Dr. Mark E. Troy, Associate Director of Data and Research Services at Texas A&M University, for his help in compiling the SET data used in this article.

References

Adams, J. B. (2005). What makes the grade? Faculty and student perceptions. *Teaching of Psychology*, *32*(1), 21–24.

Aleamoni, L. M., & Hexner, P. Z. (1980). A review of the research on student evaluation and a report on the effect of different sets of instructions on student course and instructor evaluation. *Instructional Science*, *9*(1), 67–84.

Aleamoni, L. M., & Thomas, G. S. (1980). Differential relationships of student, instructor, and course characteristics to general and specific items on a course evaluation questionnaire. *Teaching of Psychology*, *7*(4), 233.

Arreola, R. A. (2007). *Developing a comprehensive faculty evaluation system: A guide to designing, building, and operating large-scale faculty evaluation systems*. Bolton, MA: Anker.

Bailey, C. D., Gupta, S., & Schrader, R. W. (2000). Do students' judgment models of instructor effectiveness differ by course level, course content, or individual instructor? *Journal of Accounting Education*, *18*(1), 15–34.

Bennett, S. K. (1982). Student perceptions of and expectations for male and female instructors: Evidence relating to the question of gender bias in teaching evaluation. *Journal of Educational Psychology*, *74*(2), 170–179.

Blackhart, G. C., Peruche, B. M., Dewall, C. N., & Joiner Jr, T. E. (2006). Factors influencing teaching evaluations in higher education. *Teaching of Psychology*, *33*(1), 37–39.

Blunt, A. (1991). The effects of anonymity and manipulated grades on student ratings of instructors. *Community College Review*, *18*(4), 48.

Boyer, E. L. (1990). *Scholarship reconsidered: Priorities of the professoriate*. Princeton, NJ: Carnegie Foundation for the Advancement of Teaching.

Boysen, G. A. (2008). Revenge and student evaluations of teaching. *Teaching of Psychology*, *35*(3), 218–222.

Braskamp, L. A., & Ory, J. C. (1994). *Assessing faculty work: Enhancing individual and institutional performance*. San Francisco, CA: Jossey-Bass.

Brightman, H. J. (2005). Mentoring faculty to improve teaching and student learning. *Decision Sciences Journal of Innovative Education*, *3*(2), 191–203.

Bunker, M. P., & Clayson, D. E. (2008). Good grades or fair grades: The impact of expected and deserved grades on the student evaluation of instruction. In L. Robinson (Ed.), *Proceedings of the Annual Conference of the Academy of Marketing Science*, Vancouver, BC, Canada.

Burns, S. M., & Ludlow, L. H. (2005). Understanding student evaluations of teaching quality: The contributions of class attendance. *Journal of Personnel Evaluation in Education*, *18*(2), 127–138.

Cashin, W. E. (1990). *Student ratings of teaching: Recommendations for use*. Manhattan, KS: Center for Faculty Evaluation & Development, Kansas State University.

Cattell, Raymond B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245–276.

Centra, J. A. (1982). *Determining faculty effectiveness* (2nd ed.). San Francisco, CA: Jossey-Bass.

Centra, J. A. (1983). Research productivity and teaching effectiveness. *Research in Higher Education*, *18*(4), 379–389.

Centra, J. A., & Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching? *Journal of Higher Education*, *71*(1), 17–33.

Chapman, L., & Ludlow, L. (2010). Can downsizing college class sizes augment student outcomes? An investigation of the effects of class size on student learning. *Journal of General Education*, *59*(2), 105–123.

Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn? *Journal of Marketing Education*, *31*(1), 16–30.

Clayson, D. E., Frost, T. F., & Sheffet, M. J. (2006). Grades and the student evaluation of instruction: A test of the reciprocity effect. *Academy of Management Learning & Education*, *5*(1), 52–65.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research*, *51*(3), 281–309.

Cohen, P. A. (1989). Do grades influence students' evaluations of clinical courses? *Journal of Dental Education*, *53*(4), 238–240.

Divoky, J. J., & Rothermel, M. A. (1988). Student perceptions of the relative importance of dimensions of teaching performance across a type of class. *Educational Research Quarterly*, *12*(3), 40–45.

Donaldson, J. F., Flannery, D., & Ross-Gordon, J. (1993). A triangulated study comparing adult college students' perceptions of effective teaching with those of traditional students. *Continuing Higher Education Review*, *57*(3), 147–165.

Dresel, M., & Rindermann, H. (2011). Counseling university instructors based on student evaluations of their teaching effectiveness: A multilevel test of its effectiveness under consideration of bias and unfairness variables. *Research in Higher Education*, *52*, 717–737.

DuCette, J., & Kenney, J. (1982). Do grading standards affect student evaluations of teaching? Some new evidence on an old question. *Journal of Educational Psychology*, *74*(3), 308–314.

El Ansari, W., & Oskrochi, R. (2006). What matters most? Predictors of student satisfaction in public health educational courses. *Public Health*, *120*(5), 462–473.

Ellis, L., Burke, D. M., Lomire, P., & Mccormack, D. R. (2003). Student grades and average ratings of instructional quality: The need for adjustment. *Journal of Educational Research*, *97*(1), 35–40.

Elmore, P. B., & Lapointe, K. A. (1975). Effect of teacher sex, student sex, and teacher warmth on the evaluation of college instructors. *Journal of Educational Psychology*, *67*(3), 368–374.

Feldman, K. A. (1987). Research productivity and scholarly accomplishment of college teachers as related to their instructional effectiveness: A review and exploration. *Research in Higher Education*, *26*(3), 227–298.

Feldman, K. A. (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*(2), 151–211.

Feldman, K. A. (2007). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 93–143). Dordrech, the Netherlands: Springer.

Franklin, J. L., Theall, M., & Ludlow, L. (1991). *Grade inflation and student ratings: A closer look*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Gage, N. L. (1961). The appraisal of college teaching: An analysis of ends and means. *Journal of Higher Education*, *32*(1), 17–22.

Goldberg, G., & Callahan, J. (1991). Objectivity of student evaluations of instructors. *Journal of Education for Business*, *66*(6), 377.

Hippensteel, S. P., & Martin, W. (2005). Increasing the significance of course evaluations in large-enrollment geoscience classes. *Journal of Geoscience Education*, *53*(2), 158–165.

Holmes, D. S. (1972). Effects of grades and disconfirmed grade expectancies on students' evaluations of their instructor. *Journal of Educational Psychology*, *63*(2), 130–133.

Howard, G. S., & Maxwell, S. E. (1980). Correlation between student satisfaction and grades: A case of mistaken causation? *Journal of Educational Psychology*, *72*(6), 810–820.

Howard, G. S., & Maxwell, S. E. (1982). Do grades contaminate student evaluations of instruction? *Research in Higher Education*, *16*(2), 175–188.

Isely, P., & Singh, H. (2005). Do higher grades lead to favorable student evaluations? *Journal of Economic Education*, *36*(1), 29–42.

Johnson, V. E. (2003). *Grade inflation: A crisis in college education*. New York, NY: Springer.

Kaiser, Henry F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*(1), 141–151.

Kapel, D. E. (1974). Assessment of a conceptually based instructor evaluation form. *Research in Higher Education*, *2*(1), 1–24.

Kherfi, S. (2011). Whose opinion is it anyway? Determinants of participation in student evaluation of teaching. *Journal of Economic Education*, *42*(1), 19–30.

Kidd, R. S., & Latif, D. A. (2004). Student evaluations: Are they valid measures of course effectiveness? *American Journal of Pharmaceutical Education*, *68*(3), 1–5.

Langbein, L. (2008). Management by results: Student evaluation of faculty teaching and the mis-measurement of performance. *Economics of Education Review*, *27*(4), 417–428.

Lin, W. Y. (1992). Is class size bias to student ratings of university faculty. *Chinese University of Education Journal*, *20*(1), 49–53.

Lovell, G. D., & Haner, C. F. (1955). Forced-choice applied to college faculty rating. *Educational and Psychological Measurement*, *15*(3), 291–304.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential baises, and utility. *Journal of Educational Psychology*, *76*(5), 707–754.

Marsh, H. W. (2007). Do university teachers become more effective with experience? A multilevel growth model of students' evaluations of teaching over 13 years. *Journal of Educational Psychology*, *99*(4), 775–790.

Marsh, H. W., Overall, J. U., & Kesler, S. P. (1979). Class size, students' evaluations, and instructional effectiveness. *American Educational Research Journal*, *16*(1), 57–70.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist*, *52*(11), 1187–1197.

Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias, validity, or innocent bystanders? *Journal of Educational Psychology*, *92*(1), 202–228.

Mateo, M. A., & Fernandez, J. (1996). Incidence of class size on the evaluation of university teaching quality. *Educational and Psychological Measurement*, *56*(5), 771–778.

McPherson, M. A. (2006). Determinants of how students evaluate teachers. *Journal of Economic Education*, *37*(1), 3–20.

McPherson, M. A., & Jewell, R. T. (2007). Leveling the playing field: Should student evaluation scores be adjusted? *Social Science Quarterly*, *88*(3), 868–881.

McPherson, M. A., Jewell, T. R., & Kim, M. (2009). What determines student evaluation scores? A random effects analysis of undergraduate economics classes. *Eastern Economic Journal*, *35*(1), 37–51.

Morgan, J., & Davies, T. (2006). Analysis of bias in student evaluations of faculty at an all female Arab university in the Middle East. *Learning and Teaching in Higher Education: Gulf Perspectives*, *3*(2), 1–20.

Nimmer, J. G., & Stone, E. F. (1991). Effects of grading practices and time of rating on student rating of faculty performance and student learning. *Research in Higher Education*, *32*(2), 195–215.

Patrick, C. L. (2011). Student evaluations of teaching: Effects of the big five personality traits, grades and the validity hypothesis. *Assessment and Evaluation in Higher Education*, *36*(2), 239–249.

Petchers, M. K., & Chow, J. 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.

Peterson, R. L., Berenson, M. L., Misra, R. B., & Radosevich, D. J. (2008). An evaluation of factors regarding students' assessment of faculty in a business school. *Decision Sciences Journal of Innovative Education*, *6*(2), 375–402.

Pohlmann, J. T. (1975). A multivariate analysis of selected class characteristic and student rating of instruction. *Multivariate Behavioral Research*, *10*(1), 81.

Ragan, J., & Walia, B. (2010). Differences in student evaluations of principles and other economics courses and the allocation of faculty across courses. *Journal of Economic Education*, *41*(4), 335–352.

Ratz, H. C. (1975). Factors in the evaluation of instructors by students. *IEEE Transactions on Education*, *E-18*(3), 122–127.

Remedios, R., & Lieberman, D. A. (2008). I liked your course because you taught me well: The influence of grades, workload, expectations and goals on students' evaluations of teaching. *British Educational Research Journal*, *34*(1), 91–115.

Rodabaugh, R. C., & Kravitz, D. A. (1994). Effects of procedural fairness on student judgments of professors. *Journal on Excellence in College Teaching*, *5*(2), 67–83.

Sailor, P., Worthen, B. R., & Shin, E.-H. (1997). Class level as a possible mediator of the relationship between grades and student ratings of teaching. *Assessment & Evaluation in Higher Education*, *22*(3), 261–268.

Scherr, F. C., & Scherr, S. S. (1990). Bias in student evaluation of teacher effectiveness. *Journal of Education for Business*, *65*(8), 356–358.

Schuckman, H. (1990). Students' perceptions of faculty and graduate students as classroom teachers. *Teaching of Psychology*, *17*(3), 162.

Spooren, P. (2010). On the credibility of the judge: A cross-classified multilevel analysis on students' evaluation of teaching. *Studies in Educational Evaluation*, *36*(4), 121–131.

Statham, A., Richardson, L., & Cook, J. A. (1991). *Gender and university teaching: A negotiated difference*. Albany, NY: State University of New York Press.

Trick, L. R., Lehmkuhle, S., Myers, R., Graham, J., & Davis, S. L. (1993). Do grades affect faculty teaching evaluations? *Journal of Optometric Education*, *18*(3), 88–92.

Troy, M. E., Friedrich, K., & Troy, M. F. (1992). *Teaching or research? The relationship between scholarly productivity and students' judgments of teaching*. Paper presented at the annual meeting of the American Evaluation Association, Seattle, WA.

Watkins, D. (1990). Student ratings of tertiary courses for 'alternative calendar' purposes. *Assessment & Evaluation in Higher Education*, *15*(1), 12–21.

Westerlund, J. (2008). Class size and student evaluations in Sweden. *Education Economics*, *16*(1), 19–28.

Wood, K., Linsky, A. S., & Straus, M. A. (1974). Class size and student evaluations of faculty. *Journal of Higher Education*, *45*(7), 524–534.

Yeo, I.-K., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, *87*(4), 954–959.

Zabaleta, F. (2007). The use and misuse of student evaluations of teaching. *Teaching in Higher Education*, *12*(1), 55–76.

Authors

Michael D. Johnson is an assistant professor in the Department of Engineering Technology and Industrial Distribution, Texas A&M University, 3367 TAMU, College Station, TX, 77843; mdjohnson@tamu.edu.

Arunachalam Narayanan is an assistant professor in the Department of Decision and Information Sciences at the C.T. Bauer College of Business, University of Houston, 334 Melcher Hall, Houston, TX 77204-6021; anarayanan@bauer.uh.edu.

William J. Sawaya, III is an assistant professor in the Department of Management, Bowling Green State University, 3025 BAA, Bowling Green, OH, 43403; wsawaya@bgsu.edu.