Teaching, Learning, and Achievement: Are Course Evaluations Valid Measures of Instructional Quality at the University of Oregon?

Kenneth Ancell
Emily Wu

Presented to the Department of Economics at the University of Oregon, in partial fulfillment of the requirements for honors in Economics

June 9, 2017

Under the supervision of Bill Harbaugh
University of Oregon

Abstract

This study explores the legitimacy of the use of Student Evaluations of Teaching (SETs) as a measure of teaching quality. To do so, we seek to answer two questions surrounding the creation and implications of SETs. Using data from the University of Oregon (UO) we first analyze the influence of a variety of factors commonly hypothesized to bias SET scores. Second, we investigate the relationship between SET scores and future student achievement. We find that a many of these factors influence SET scores, and that SET scores for a class are not a useful measure for predicting how well students will do in future classes. These findings suggest that SET scores are not a valid measure of teaching quality at the UO.

Table of Contents

Introduction

In many institutions of higher learning, Student Evaluations of Teaching (SETs) are used as a tool for students to evaluate their instructor's performance. Though not standardized across institutions, SETs typically feature questions about a variety of instructor and course characteristics, such as the overall quality of the course and the overall quality of the instructor's teaching. Students are asked to answer these questions according to a set scale and their answers are converted into a numerical score.

The results of these evaluations serve a variety of purposes within the institution. Their primary function is as a mechanism through which instructors can improve their teaching. However, these scores are also often incorporated into the decision making process of awarding tenure, teaching awards, and merit increases. In addition, students frequently utilize previous terms' evaluation scores in selecting classes.

Despite the wide-ranging implications of these evaluations, there exists a substantial collection of evidence suggesting that SETs are not a valid measure of teaching quality. Instead, this evidence suggests that SETs are influenced by a variety of factors irrelevant to an instructor's actual teaching ability. Although these factors, which include elements such as instructor gender and race, should not influence an instructor's ability to teach effectively, some research suggests that they do influence an instructor's SET scores. Additionally, there are other factors, such as class size and class level, that may influence SET scores but also may not be accounted for when these scores are used to evaluate teaching quality. Thus, SET scores may not truly reflect an instructor's teaching quality and this disparity between quality and score may negatively impact an instructor's outcomes in decisions like tenure status.

Literature Review

There is an expansive base of literature concerning the use of SETs that dates back nearly 90 years, beginning with Herman H. Remmers and G. C. Brandenburg's *Experimental Data on the Purdue Rating Scale For Instructors* in 1927. A significant portion of this literature is dedicated to exploring the impact of a variety of factors perceived to bias SET scores. One such factor is grades. The prevailing hypothesis concerning the relationship between grades and SET scores posits that awarding higher grades will lead to better evaluation scores, as students are more likely to have favorable attitudes regarding their instructors if they are more satisfied with their grade. Herbert W. Marsh and Lawrence A. Roche (2000) term this hypothesis the grading-leniency hypothesis. They state "the grading-leniency hypothesis proposes that instructors who give higher-than-deserved grades will be rewarded with higher-than-deserved SETs, which constitutes a serious bias to SETs."[1]

According to Kenneth A. Feldman (2007), "almost all of the available research does show a small or even modest positive association between grades and evaluation (usually a correlation somewhere between +.10 and +.30)."[2] One study that supports this correlation between grades and SET scores is Anthony G. Greenwald and Gerald M. Gillmore's *Grading Leniency is a Removable Contaminant of Student Ratings* (1997). In this paper, Greenwald and Gillmore conclude that "in the population of courses included in the University of Washington data sets, changing from giving grades one standard deviation below the university mean to one standard deviation above should produce a one standard deviation change in one's percentile

[1] Marsh and Roche 2000, 1191
[2] Feldman 2007, 99
[3] Feldman 2007, 99; Greenwald and Gillmore 1997, 1214
[4] Langbien 2008, 419
[5] Langbien 2008, 419
[6] Marsh 1987, 21
[7] Marsh 1987, 21
[8] Marsh & Roche 2000, 1191

rank in the university's student ratings"[3] This type of correlation is also found by Laura Langbein (2008) who writes "actual and expected grades both have a significant, positive effect on SETs, controlling for faculty *or* course fixed effects, or for faculty *and* course fixed effects."[4] Langbien also details the detrimental effects of this relationship between grades and SET scores when she states:

> The overall implication is that students, administrators and faculty are engaged in an individually rational but arguably socially destructive game. Administrators want higher SETs because it leads to higher grades and higher student retention rates, which means more tuition and tax revenues. Faculty want higher teaching evaluations because it leads to higher salaries, and students want higher grades for the same reason. But the overall social effect is to make both the SET a faulty signal of teaching quality and grades a faulty signal of future performance on the job. No student, no individual faculty member, no individual college or university administrator, and no college or university institution have much of an incentive to break this vicious cycle[5]

Based on the grading-leniency hypothesis, a causal relationship between grades and SET scores would clearly have serious ramifications across the higher education landscape. If instructors were able to manipulate their SET scores by simply awarding higher grades, these scores would lose any value as a measure of teaching quality. Given the variety of uses that depend on SET scores being an accurate and valid measure of the quality of an instructors teaching, the potential that these ratings are biased by grades is concerning.

However, scholars have posited alternative hypotheses to the grading-leniency hypothesis. One such hypothesis is the validity hypothesis. Marsh explains that "the 'validity hypothesis' proposes that better expected grades reflect better student learning, and that a positive correlation between student learning and student ratings supports the validity of student

[3] Greenwald & Gillmore 1997, 1214
[4] Langbien 2008, 419
[5] Langbien 2008, 419

ratings."[6] If this hypothesis truly explained the correlation between grades and SET scores, the relationship would be far less troubling. In this case, students would learn more in classes that are taught by better teachers and this learning would be reflected in higher grades, while the quality of instruction that led to those grades are reflected in SET scores.

Marsh also proposes the student characteristics hypothesis as an explanation for the relationship between grades and SET scores. This hypothesis "proposes that pre-existing student characteristics may affect student learning, student grades, and teaching effectiveness so that the expected grade effect can be explained in terms of other variables."[7] Marsh and Roche suggest prior subject interest as one example of a pre-existing student characteristic that could contribute to the grade-SET score relationship.[8] Once again, this hypothesis has far milder consequences than the grading leniency hypothesis. If a student is more interested in a particular subject, they may be more engaged and receptive in the classroom, which can then lead to more learning and better grades. Thus, regardless of anything an instructor does, the strength of the correlation between grades and SET scores may vary due to pre-existing student characteristics.

Given the variety of hypotheses and the ramifications of them, it is difficult to draw conclusions from results showing a correlation between grades and SET scores. Nonetheless, Marsh and Michael J. Dunkin (1992) write:

> Evidence from a variety of different types of research clearly supports the validity hypothesis and the student characteristics hypothesis, but does not rule out the possibility that a grading leniency effect operates simultaneously. Support for the grading leniency effect was found with some experimental studies, but these effects were typically weak and inconsistent, may not generalize to nonexperimental settings where SETs [students' evaluations of teaching effectiveness] are actually used, and in some instances may be due to the violation of grade expectations that students had falsely been led to expect or that were applied to other students in the same course. Consequently, while it is possible

[6] Marsh 1987, 21
[7] Marsh 1987, 21
[8] Marsh & Roche 2000, 1191

> that a grading leniency effect may produce some bias in SETs, support for this suggestion is weak and the size of such an effect is likely to be insubstantial in the actual use of SETs[9]

While it may be unlikely, as Marsh and Dunkin assert, that grades can lead to bias in SET scores, existing research does not rule out the possibility that there is in fact bias or that this bias is substantial enough to corrupt SETs as they exist today. Thus, further exploration into the relationship between grades and SET scores is warranted.

Another factor that has been suggested as a potential source of bias within SETs is gender. Given the variety of uses for SETs within the world of higher education, a systematic gender bias within SET scores would be significant contributor to inequality between male and female instructors. As a result of the gravity of these implications, researchers have attempted to unmask the existence of gender bias within SET scores. Thus far, the findings concerning the impact of gender on SET scores have been mixed. In their paper *Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness*, Anne Boring, Kellie Ottoboni, and Philip B. Stark (2016) find that "average SET are significantly associated with instructor gender, with male instructors getting higher ratings (overall *p*-value 0.00). Male instructors get higher SET on average in every discipline… with two-sided *p*-values ranging from 0.08 for history to 0.63 for political science."[10] Daniel S. Hamermesh and Amy Parker find similar results, noting "significantly lower [ratings] received by female instructors, an effect that implies reductions in average class ratings of nearly one-half standard deviation." [11] These results clearly support the hypothesis that SET scores are biased against female instructors.

However, as Lillian MacNell, Adam Driscoll, and Andrea N. Hunt (2015) explain, "it is

[9] Marsh & Dunkin 1992, 202
[10] Boring, Ottoboni, and Stark 2016, 6
[11] Hamermesh and Parker 2005, 373

difficult to separate the gender of an instructor from their teaching practices in a face-to-face classroom."[12] To account for this difficulty, MacNell, Driscoll, and Hunt conducted an experiment using online courses where two instructors (one male and one female) each taught 2 sections of a class, but taught one section under the identity of the other instructor. Thus, "if gender bias was present, than the students from the two groups who believed they had a female instructor should have given their instructor significantly lower evaluations than the two groups who believed they had a male assistant instructor."[13] Based on this experimental design, MacNell, Driscoll, and Hunt's results "support the existence of gender bias in that students rated the instructors they perceived to be female lower than those they perceive to be male, regardless of teaching quality or actual gender of the instructor."[14]

Although, Boring, Ottoboni, and Stark, Hamermesh and Parker, and MacNell, Driscoll, and Hunt all find evidence indicating that gender bias exists in SET scores, numerous other studies have failed to find this type of evidence. For example, in their study of SETs, Patricia B. Elmore and Karen A. LaPointe (1974) found that "in general, there seemed to be few meaningful differences between male and female faculty."[15] These findings are echoed by Francisco Zabaleta (2007), who concluded that "gender does not play a significant role in either evaluations or grades."[16] In fact, Zabaleta found that "female instructors received slightly better evaluations (4.13 against 4.10) and they assigned better grades (2.95 against 2.92) but the differences were minimal."[17] In addition to individual studies that did not find evidence suggesting that gender bias exists in SETs, in his evaluation of existing literature on the

[12] MacNell, Driscoll, and Hunt 2015, 292
[13] MacNell, Driscoll, and Hunt 2015, 292
[14] MacNell, Driscoll, and Hunt 2015, 300
[15] Elmore and LaPointe 1974, 387
[16] Zabaleta 2007, 59
[17] Zabaleta 2007, 58

relationship between gender and SET scores, Feldman stated "much of the relevant research has *not* found any differences between men and women teachers in either their students' global or specific evaluations of them to begin with."[18] Furthermore, Feldman noted "in those studies in which statistically significant differences were found, more of them favored women than men."[19]

Although Feldman asserts that the existing literature represents a consensus among researchers that gender does not impact SET scores in a manner that suggests that the mechanism is biased against female instructors, given the many well known biases against women in academia and in professional life in general, the existence of studies finding evidence suggesting that SETs are biased against women is cause for concern.

Academic rank has also been hypothesized as a potential source of bias within SETs. Within the literature, some have found that instructors of a higher rank receive higher SET scores. Feldman points to studies by Centra and Creech (1976), Brandenburg and Aleamoni (1976), and Brandenburg, Slinde, and Batista (1977) as examples of research that has found no differences in ratings among faculty members (full professors, associate professors, assistant professors, and instructors) but have found that "each of these four groups of teachers was somewhat more highly rated than was the group of graduate teaching assistants included in the study."[20] This finding is in line with the prevailing sentiment regarding the relationship between academic rank and teaching quality. As Feldman explains, "at certain colleges and universities teachers of higher rank may in fact typically be somewhat better teachers and thus "deserve" the slightly higher ratings they receive."[21] Since teaching is one component considered when evaluating tenure and promotion, it serves to reason that individuals who have reached higher

[18] Feldman 1993, 46
[19] Feldman 2007, 97
[20] Feldman 1983, 6
[21] Feldman 2007, 98

ranks would be higher quality teachers.

However, some studies have found no relationship between academic rank and SET scores. According to Arnold S. Linsky and Murray A. Straus (1975) "academic rank is uncorrelated with overall teaching score (r=.00, N=3530)."[22] Dorthoy D. Nevill, William B. Ware, and Albert B. Smith (1978) also reach similar conclusions. They write, "students appear to rate teaching assistants and faculty members in a similar fashion, both in terms of the ratings themselves and the conceptual framework within which these decisions are made."[23] Lawrence M. Aleamoni (1987) also finds no significant relationship between academic rank and SET scores in a variety of studies including Aleamoni and Graham (1974), Aleamoni and Thomas (1980), and Aleamoni and Yimer (1973).[24]

Although many studies find evidence that more highly ranked instructors receive higher SET scores and many others find no evidence of any relationship between rank and SET scores, still some others find that lower ranked instructors receive higher SET scores. For example, A. Paul King (1971) reports "it appears from this study, that students rated those instructors higher who… [had] a professional rank lower than a professor."[25] Hamermesh and Parker also report findings of an inverse relationship between academic rank and SET scores. They write, "non-tenure-track instructors receive course ratings that are surprisingly almost significantly higher than those of tenure-track faculty."[26] To explain this relationship, Hamermesh and Parker write "this may arise because they are chiefly people who specialize in teaching rather than combining teaching and research, or perhaps from the incentives (in terms of reappointment and salary) that

[22] Linsky & Straus 1975, 99
[23] Nevill, Ware, and Smith 1978, 36
[24] Aleamoni 1987, 114
[25] King 1971, 48
[26] Hamermesh and Parker 2005, 373

they face to please their students."[27]

David N. Figlio, Morton O. Schapiro, and Kevin B. Soter (2015) find evidence that suggests that Hamermesh and Parker's findings that non-tenure-track instructors receive better SET scores is the result of genuinely higher teaching quality. In their study, Figlio, Schapiro, and Soter utilize data on Northwestern University freshman to explore whether or not these students learn more from tenure track or non-tenure-track faculty members. Ultimately, they conclude that "contingent faculty at Northwestern University not only induce first term students to take more classes in a given subject than do tenure line professors, but also lead the students to do better in subsequent course work than do their tenure track/tenured colleagues."[28] These findings suggest that a positive correlation between non-tenure-track faculty status and SET scores is a valid relationship, as this class of instructors actually inspires higher future student achievement and thus can be considered higher quality teachers.

While the consequences of a bias to SET scores due to differences in instructors' academic rank may be less dire than those of a bias due to grades or gender, the conflicting evidence surrounding the relationship between academic rank and SET scores suggests that this bias may in fact exist. Due to these confounding findings, further exploration of this relationship is warranted.

In addition to grades, gender, and academic rank, some studies have reported that the subject an instructor teaches can impact their SET scores. For example, Tanya Beran and Claudio Violato (2005) find that "courses in social sciences received significantly higher ratings than courses in natural sciences."[29] Similarly, Edward L. Delaney Jr. (1976) asserts that "it is

[27] Hamermesh and Parker 2005, 373
[28] Figlio, Schapiro, and Soter, 2015, 723
[29] Beran and Violato 2005, 599

noteworthy to observe that the beta weights for the more codified fields, such as biology, psychology, health sciences, mathematics, physical science and business, seemed to increase in negative values, predicting lower ratings."[30] Most recently, Bob Uttl and Dylan Smibert (2017) write that "Math classes received much lower average class summary ratings than English, History, Psychology or even all other classes combined, replicating previous findings showing that quantitative vs. non-quantitative classes receive lower SET ratings."[31] Additionally, Uttl and Smibert note that "whereas the SET distributions for non-quantitative courses show a typical negative skew and high mean ratings, the SET distributions for quantitative courses are less skewed, nearly normal, and have substantially lower ratings."[32]

In addition to reporting their findings concerning the relationship between subject and SET scores, Uttl and Smibert also address the potential ramifications of this relationship. They find that "professors teaching quantitative courses are far less likely to be tenured, promoted, and/or given merit pay when their class summary ratings are evaluated against common standards, that is when the field one is assigned to teach is disregarded. They are also far less likely to receive teaching awards based on their class summary SET ratings."[33] While these findings do not necessarily indicate that SETs are biased due to subject, the existence of this type of bias would create a variety of problems given how SETs are currently utilized in college and university settings. For one, SETs would not be able to be used to compare professors within departments, as those who teach quantitative courses would receive artificially lower SET scores than their peers. Also, SETs would no longer be a reliable medium for comparing instructors across departments in the process of awarding teaching awards. Thus, it is important to

[30] Delaney 1976, 11
[31] Uttl and Smibert 2017, 9
[32] Uttl and Smibert 2017, 9
[33] Uttl and Smibert 2017, 9

determine if the differences in SET scores across subject are due to a bias to SETs or due to some other factor.

One hypothesized potential source of bias to SETs that has not received as much attention in the literature is race. While a bias against certain races within SETs would create legal, ethical, and practical problems in the use of SETs, many studies have not included an exploration of the role of race in SET scores in their analysis. However, a select few studies have focused on the effect of race on SET scores. One such study is Bettye P. Smith's *Student Ratings of Teaching Effectiveness: An Analysis of End-of-Course Faculty Evaluations* (2007). In this study, Smith explores the ratings that faculty members of different races receive on a variety of questions included on a SET questionnaire. In her analysis, Smith finds that "White faculty had significantly higher mean scores than Black faculty on the composite of multidimensional items and the two global items, *overall value of course* and *overall teaching ability.*"[34] As Smith notes "the findings from this study are significant because they provide empirical data about student evaluations of Black faculty and contribute to the dialogue about the use of student end-of-course evaluations in making decisions about promotion, tenure, merit increases, and teaching awards."[35]

In order to further explore this issue of potential bias in students' perceptions of their instructors, Kristin J. Anderson and Gabriel Smith (2005) conducted an experiment where they created a syllabus for a class then altered the syllabus to have different teaching styles, genders, and ethnicities. Then they asked students to rate the hypothetical course and instructors on a variety of factors including warmth, availability, knowledge of the topic, preparedness and capability, and lack of objectivity and political bias. In their analysis Anderson and Smith found

[34] Smith 2007, 796
[35] Smith 2007, 798

that "Anglo women professors with strict teaching styles were viewed as warmer than Latina professors with the same teaching style."[36] Additionally, Anderson and Smith write that

> the flip side of this pattern also seems to be true: Latino professors with lenient teaching styles, particularly Latinas, were rated as warmer than Anglo professors with the same teaching styles (nonsignificant trend). Therefore, ratings of professor warmth and availability for Latino professors appear to be contingent on their teaching style, whereas the rating of Anglo professors' warmth is less contingent on teaching style. Thus, this pattern seems to reveal a double standard in the evaluation of Latino and Anglo professors[37]

Similarly to these results found by Smith and Anderson and Smith, Hamermesh and Parker also find evidence that instructors of certain races may receive worse SET scores. In their paper, they conclude that "minority faculty members receive lower teaching evaluations than do majority instructors and non-native English speakers receive substantially lower ratings than do natives."[38] Given these results and the relative lack of exploration into this phenomenon that currently exists in the literature on SETs, further exploration into this topic is surely warranted.

In addition to the aforementioned effects regarding grades, gender, academic rank, subject, and race, various studies within the literature have reported other assorted factors that are believed to potentially influence SET scores. For example, Hamermesh and Parker found that an instructor's beauty impacts their SET scores. They explain "the effects of differences in beauty on the average course rating are not small: Moving from one standard deviation below the mean to one standard deviation above leads to an increase in the average class rating of 0.46, close to a one-standard deviation increase in the average class rating."[39] Another factor, reported by Michael A. McPherson, R. Todd Jewell, and Myungsup Kim (2009), is age. In their study,

---

[36] Anderson and Smith 2005, 193
[37] Anderson and Smith 2005, 196
[38] Hamermesh and Parker 2005, 373
[39] Hamermesh and Parker 2005, 372

they state "additional years of instructor *age* lead to a worsening of evaluation scores."[40] Additionally, in his study *Research Productivity and Teaching Effectiveness,* John Centra (1981) writes "student ratings of teaching, as the present study and others have demonstrated, are also unrelated or only modestly related to research productivity."[41] This finding is contrary to the prevailing sentiment regarding the relationship between research productivity and teaching quality, which Centra describes when he states "the belief that teaching and research performance are related is undoubtedly stronger than this or any other study has shown. When peers were asked to judge their colleagues' professional performance, their ratings of teaching and research effectiveness correlated with each other (Wood, 1978)."[42]

In addition to research on these factors, Feldman surveys the literature relating to a variety of course characteristics that have been theorized as potential influencers of SET scores. One such course characteristic is the "electivity" of a course. In exploring the impact of electivity on SET scores he writes "the relationship between the percentage of students taking a course as an elective (that is the "electivity" of the class for students in it) and the ratings of the teacher and the course is generally positive and of small to moderate strength."[43] Feldman also examines the literature concerning the relationship between course level and student ratings. In doing so Feldman writes that "the positive association between course level and ratings is clear and relatively consistent across various rating items" but notes that this relationship "does tend to be quite weak in strength" and that "a positive association between the two variables under consideration is not universally found."[44] In addition to electivity and course level, Feldman also

[40] McPherson, Jewell, and Kim 2009, 45
[41] Centra 1981, 11
[42] Centra 1981, 10
[43] Feldman 1978, 218
[44] Feldman 1978, 216

evaluates existing literature concerning the connection between class size and SET score. He finds that "about one third of these studies find essentially no relationship between size and ratings" and that "the rest (roughly two thirds) of these correlational analyses find indications of a negative relationship—the smaller the size of the class, the higher the ratings."[45] Finally, Feldman also investigated the effect of class meeting time on SET scores. He states "it might be thought that students' general preferences for some class times rather than others might 'spill over' into their ratings of courses and the instructors themselves" but concludes that "little support for this notion exists."[46]

Clearly, many potential influencers of SET scores exist. Grades, instructor gender, academic rank, subject, instructor race, electivity, course level, class size, and class meeting time have all been hypothesized and explored as potential sources of bias within SET scores. The results of these explorations are mixed. On some factors, such as class meeting time, previous research has formed a consensus on their true impact. On other factors, such as grades, scholars remain divided on the true effect. The existing literature on these factors suggests that further exploration of the role they play in SETs is necessary.

Although substantial evidence suggests that many factors have varying impacts on SET scores, many researchers have concluded that SETs are valid measures of teaching quality. To evaluate if SETs are valid measures of teaching quality, one must first establish what quality teaching actually is. Unfortunately, as Dennis C. Clayson (2009) explains, "no one has given a widely accepted definition of what 'good' teaching actually is, nor has a universally agreeable criterion of teaching effectiveness been established."[47] However, Clayson notes, "both defenders

[45] Feldman 1978, 206
[46] Feldman 1978, 219
[47] Clayson 2009, 16

and detractors of SET generally agree that students will learn more from good teachers."[48] Thus, "if the process is valid, then there should be an association between student learning and the evaluations that students give of classes and instructors."[49] Given this connection, many researchers have evaluated the connection between evaluation scores and student learning in order to determine the validity of SETs.

Two such researchers are Richard John Stapleton and Gene Murkison (2001). In their study, they utilize a SET question concerning the amount learned in the course to measure student learning. They then explore the correlation between this question and a question concerning instructor excellence. Ultimately, Stapleton and Murkison find that "student evaluations are generally valid by showing a positive relationship between instructor excellence scores and learning produced in the course."[50] While Stapleton and Murkison utilize a SET question to measure student learning, Trinidad Beleche, David Fairris, and Mindy Marks (2012), employ a different method. Their measure of student learning comes from "a unique setting in which students take a pre-test placement exam and post-test exit exam, which is common to all students and is graded by a team of instructors instead of the instructor of record for this course."[51] In utilizing this measure, Beleche, Fairris, and Marks find:

> in specifications that use the common post-test as a measure of learning, there is a consistently positive and statistically significant relationship between individual student learning and course evaluations. The main relationship between learning and course evaluations is strengthened by ability controls and is robust to the inclusion of instructor and section fixed effects. While the estimated relationship is positive and statistically significant, the quantitative association is not large in magnitude, suggesting that it may be prudent for institutions wishing to capture the extent of knowledge transmission in the classroom to explore measures beyond student course evaluations[52]

---

[48] Clayson 2009, 17
[49] Clayson 2009, 17
[50] Stapleton and Murkison 2001, 279-80
[51] Beleche, Fairris, and Marks 2012, 718
[52] Bleeche, Fairris, and Marks 2012, 718

Finally, in his review of existing research addressing the validity of SETs, Marsh concludes that "SETs are multidimensional, reliable and stable, primarily a function of the instructor who teaches a course rather than the course that is taught, relatively valid against a variety of indicators of effective teaching, relatively unaffected by a variety of potential biases, and seen to be useful by faculty, students, and administrators."[53]

Ultimately, the existing literature concerning SETs paints a complicated picture. On one hand, contradictory evidence exists concerning the influence of a variety of factors on SET scores. On the other hand, many researchers have still concluded that SETs are valid measures of teaching quality. Additionally, it is difficult to determine the root causes of the differences in findings. Differences in methodologies, samples, and course evaluation questions are all plausible sources of the incongruity of these results. Given this dissonance, this research will seek to evaluate both the impact of these various factors on SET scores and the relationship between SET scores and student learning at the University of Oregon. We hope to conclude whether or not the University of Oregon's course evaluation system, as it is currently constructed, is a valid measure of teaching quality. Because we use multiple methods to investigate the validity of SET scores rather than relying on a single indicator, our research will bring new insights into the existing literature. These methods allow us to incorporate several distinct and contradictory relationships into a comprehensive investigation of SETs and their relationship with learning outcomes.

Data

Our data set is a composition of four separate sources from the University of Oregon, a large (20,000+), public, 4-year university. The data spans from from Fall 2010 to Spring 2016

[53] Marsh 2007, 372

and includes undergraduate courses in the colleges of Architecture and Allied Arts, Business, Education, Humanities, Journalism, Law, Music and Dance, Natural Sciences and Social Sciences. Physical Education courses were dropped for the purpose of this study.

Our first data source contained information regarding course evaluation scores. In 2007, the University of Oregon began administering its course evaluations online through its student portal DuckWeb.[54] The course evaluation process begins at midnight on the Friday before the University's dead week and closes early Monday morning before the final exam period begins.[55] If students do not either complete their evaluation or indicate they decline to respond, the University withholds their final term grades for two weeks.[56] The course evaluation form asks 12 standard questions. Seven of the twelve questions ask students to indicate measures of quality using a Likert Scale (1=Unsatisfactory, 2=Somewhat inadequate, 3=Adequate, 4=Good, 5=Exceptional). Of these seven questions, three of the concern characteristics of the course, three concern instructor characteristics, and the final question asks students to rate the amount they learned in the course. The next two questions provide a field for students to leave written comments regarding the course and the instructor. The final three questions ask about the percentage of time a student attends class, the amount of time outside of class they spend on the course, and their expected grade in the course. The last five questions are omitted from the course evaluation database. However, our data does contain each instructor's course evaluation score for classes that took place within our sample period, the department average on each course evaluation question, the class enrollment, the average class size in the department, and the percentage of students that completed evaluations.

[54] Office of the Registrar
[55] Office of the Registrar
[56] Office of the Registrar

Our second data source contains information on the grades awarded in classes that met the requirements of the Family Education Rights and Privacy Act (FERPA). To maintain compliance with FERPA, the University requires that for course grade data to be published, actual class enrollment must be greater than or equal to ten students, all students in the class must not receive the same grade, and the class cannot award every student the same grade except for five or fewer students. Given these conditions 67% of the data was redacted. After the redaction 36,914 observations remained, implying that, of the 79,118 classes merged from the data sets, 42,204 were redacted by the Registrar. Using the information on the grade distribution of the courses in our sample, we calculated the average GPA of each course to use in our model.

Our third data source is several years worth of published Salary Reports from the University's Office of Institutional Research. We employed the data on hiring and compensation contained within this source to determine the academic rank for the instructors in our composite data set. By matching across the various years of Salary Reports, we were able to approximate the change in a faculty member's rank across our sample period. Due to the inconsistencies in job title (eg. Senior Instructor versus Instructor), academic rank was coded into five categories for simplicity: Full Professor, Assistant Professor, Associate Professor, Instructor, and Other. Full and Associate Professor are tenured positions, Assistant Professor is tenure track, Instructor is non-tenure track, and "Other" indicates that the instructor has no faculty rank. The "Other" category is likely composed predominately of Graduate Employees (GTFs).

Our final data source is transcript data and student demographics from the Office of the Registrar. This source contains demographic and grade information on students in the Lundquist College of Business and the School of Journalism and Communications. The demographic information in this dataset includes gender, race, high school and college GPA, standardized test

scores, age, state residency status, and international student status. The grade information includes the grade a student received in a class, the class' instructor, and the term the class was taken.

One piece of data we did not receive was the race and gender of the instructors. In order to create this data, we employed the R packages gender and ethnicity, which codes gender based on first name and ethnicity based on last name. Both the R packages utilized historical data sets from the US Census Bureau to code for gender and ethnicity. For race, encoding is based on probabilities found in tabulations of surnames occurring 100 or more times in the 2010 Census returns. If the surname is not found, then the probability is coded based on the demographic breakup of a specified geographic county. The R package on gender specifies a range of birth years, and because the average age across fields for doctorate degrees awarded is 33 (see "Statistical Profile"), people with the rank of "Other" were coded with birth years from the ages of 22 to 33 and the other instructors were coded for the ages of 33 to 63 (the average age of retirement). The R package automatically assigned an individual a gender if their probability of being one gender or the other was over 0.5. However, given the various complications in assigning race, we chose to err on the side of caution and only assign a race if the probability of being a given race was greater than 0.9 (see table 7).

Methodology

Our first model, which we refer to as the SET Score Model, seeks to explore the relationship between various factors hypothesized in the literature as potential sources of bias within SETs. To this end we specified a regression where the dependent variable is course evaluation scores. We treat these scores as a function of instructor characteristics, course characteristics, and of the class GPA. The instructor characteristics include an instructor's

gender, race, and academic rank. The course characteristics include the course level and the class size. These factors were included as a result of some indication of importance in the existing literature and of availability of data. We included the average GPA of the class to account for student achievement, as student achievement is significantly influenced by teaching quality. However, as discussed previously, the relationship between grades and SET scores is a point of contention. Ultimately, if course evaluation scores are indeed an unbiased and valid measure of teaching quality, then changes in instructor characteristics should not cause variability across the course evaluation scores, or should not drive up one specific course evaluation question's scores.

$$AvgCourseEval_j = \beta_0 + \beta_1 AvgGPA_j + \beta_2 ClassEnroll_j + \beta_3 InstructorRank_{ij} + \beta_4 Race_i + \beta_5 Gender_i + \alpha_i + \mu_{tj} + \varepsilon_{jit} \quad (1)$$

We also estimated versions of our primary regression using the following interaction variables:

$$\beta_6 AvgGradeDist_j * Gender_i$$
$$\beta_7 AvgGradeDist_j * Race_i$$

Our second model seeks to explore the relationship between course evaluations and teaching quality. The challenge presented in specifying this model is in determining an objective measure of teaching quality. There are a variety of dimensions to effective teaching, making it difficult to find a catchall measurement of teaching quality. However, as Clayson explains, it is generally agreed upon that students learn more from better teachers. Given this consensus, we decided to use future student achievement as a proxy for teaching quality. To create a student achievement metric, we used our student grade data to create pairs of prerequisite and post-requisite courses. Then we normalized students' grades in both the prerequisite and the post-requisite to the class average and took the difference of the two scores. To control for the variations in achievement between courses, we normalized relative to each class' average GPA. This decision is based on the fact that, depending on the performance of students in a class, a

grade in one class is not necessarily equivalent to that same grade in another class, even if the course is exactly the same. This process gave us our measure of future student achievement and thus our measure of teaching quality.

To evaluate the relationship between course evaluations and teaching quality we specified a regression with our measure of achievement as the dependent variable. We refer to this model as the Future Student Achievement Model. Our independent variables included student characteristics, instructor characteristics, class characteristics, and the course evaluation questions. The student characteristics included in the model are race, gender, state residency status, college GPA, SAT Math scores, SAT Verbal[57] scores, age, and international student status. The instructor characteristics included in the model are race, gender, and academic rank. The course characteristic included in the model is class size. Finally, the seven course evaluation questions concerning instructor and course quality were included in the model.

$$StudentAchievement_j = \beta_0 + \beta_1 StudentCharacteristics_j + \beta_2 InstructorCharacteristics + \beta_3 ClassSize_{ij} + \beta_4 CourseEvaluation\ Questions_{\square} + \alpha_i + \mu_{tj} + \varepsilon_{jit}$$

Additionally, we specified an additional version of this model to include the interaction variable:

$$\beta_5 InstructorRace_j * StudentRace_i$$

In both models we included fixed effects to control for variation across the subject of the course, the year the course was taught, and the term in which the course was taught.

Results – SET Score Model

Our initial regression model seeks to measure the influence of various factors on UO course evaluation scores. The results can be seen in table 1 and the summary statistics for the

[57] For those students who took the ACT rather than the SAT, we converted their ACT score to an SAT score following the guidelines set out in the ACT-SAT concordance tables

variables included in the model can be found in table 2. Based on existing literature, we have chosen to include variables that allow us to explore the impact of grades, gender, rank, race, course level, and class size on SET scores. When evaluating the effect of grades on SET scores, it is clear that a positive correlation between the two variables exists. For all seven course evaluation questions, a positive and statistically significant relationship exists between grades and SET scores. The magnitude of this relationship ranges between 0.182 and 0.319. This relationship suggests that a one point increase in the GPA of a class could lead to between a 0.182 and 0.319 point increase in the instructor's evaluation score, depending on the question of interest.

Looking at gender, a negative and statistically significant correlation exists across all course evaluation questions. However, this relationship is small in magnitude, ranging from -0.0578 to -0.0158. For example, for the question regarding instructor quality, our results suggest that being a female instructor may lead to a course evaluation score that is 0.0578 points lower than if the instructor were male.

For academic rank, the results are less clear-cut. For six out of the seven course evaluation questions, a positive and statistically significant relationship exists between being a non-tenure-track instructor and SET scores. These results are similar in magnitude to the relationship between grades and SET score. For non-tenure-track instructors, the coefficient ranges between 0.0175 and 0.0604. This implies that being a non-tenure-track instructor can lead to a SET score that is between 0.0175 and 0.0604 points higher than if the instructor was a tenure-track faculty member. Additionally, we included a variable to explore the effect of being a GTF on SET scores. The results of this exploration yielded mixed results. For five of the seven course evaluation questions, we found a negative relationship between GTF status and SET

score. Of these seven coefficients, three of them are statistically significant. The other two course evaluation questions exhibit a positive and statistically significant relationship. The magnitudes of the relationship between GTF status and SET scores range from -0.0127 and 0.0586.

Our results indicate that the relationship between being a non-white instructor and SET scores are generally inconclusive. We report coefficients ranging from -0.00191 to 0.0361. However, none of these coefficients are significant at the 95% confidence level. On the other hand, we found that the relationship between being a white instructor (with nonwhite instructors as the omitted category) and SET scores was generally negative. Our reported coefficients range between -0.0163 and -0.00187. Of the seven course evaluation questions, we found a statistically significant relationship for four of the seven questions at the 95% Confidence Level.

Similar to academic rank, the results of our evaluation of the effect of course level on SET scores are inconclusive. Three of the seven course evaluation questions exhibit a negative relationship with upper division courses. Of these three questions, only one had a statistically significant negative relationship. The other four course evaluation questions exhibited a positive and statistically significant relationship between course level and SET scores. For all seven questions, the relationship between teaching an upper division course and SET scores is between -0.00623 and 0.0401.

Finally, using our initial regression, we were able to evaluate the impact of class size on SET scores. We observe negative and statistically significant relationships between class size and SET scores for all seven course of the evaluation questions. However, similar to the coefficients reported for the white instructor variable, these relationships are extremely small in magnitude. For the six statistically significant coefficients, their magnitudes range from

-0.000784 to -0.00000938 (the magnitudes of an increase in 100 students would range from -0.0784 to -0.000938).

In addition to examining the impact of these various factors on SET scores, we also explored the impact of these same factors on the response rate of SET scores. Of the eight factors included in our model, five factors exhibit a statistically significant relationship with response rate. One such factor is grades. Our results suggest that increasing course GPA by one point may increase response rate by 3.586 percentage points. Another statistically significant relationship was between upper division courses and response rate. The implication of this result is that teaching an upper division class can lead to a response rate that is 4.526 points lower than that of a lower division classes. Additionally, we find that being a non-tenure-track instructor (-0.494 points) is negatively correlated with higher response rates while being a GTF (0.824 points) is positively correlated with higher response rates relative to being a tenure-track professor. The final statistically significant relationship we find is between class size and response rate. Similar to the effect of class size on SET scores, the relationship between class size and response rate is quite small, with a coefficient of 0.00472 (0.472 for a 100 student increase in class size). However, it remains unclear what influence increased response rate has on the accuracy of SET scores

Results – Future Student Achievement Model

This model seeks to evaluate the relationship between course evaluation questions and change in future student achievement in order to evaluate whether or not these questions are valid tools for measuring teaching quality. To do so, we included various student and course characteristics that may impact a student's achievement so as to not attribute the effect of any of

these factors to the questions themselves. The results of this model can be seen in table 4 and the summary statistics of the variables included in the model can be found in table 5.

The student characteristics in our model include their gender, race, residency status, college GPA, SAT Math score, SAT Verbal score, age, and international student status. Among these factors, only residency status displays a statistically significant relationship with future student learning. The coefficient on residency status is 0.0565 and is statistically significant at the 95% confidence level. This suggests that being an out of state student leads to an improvement in achievement relative to their class that is 0.0565 points higher than in state students. Of the other factors, only gender and college GPA also had a positive relationship with future student achievement. For race, SAT Math, SAT Verbal, age, and international student status, the relationship with future student achievement is negative. Of these results, the relationship between race and future student achievement is concerning. Our findings imply that being a non-white student can lead to an improvement in achievement relative to their class that is 0.231 points lower than white students. Another notable result is the negative relationship between SAT Math and SAT Verbal. Our findings suggest that a one point increase in SAT Math may lead to a smaller improvement in achievement relative to their class by 0.000270 points. For SAT Verbal a one point increase in SAT Verbal can lead to a smaller improvement in achievement relative to their class of 0.000174 points. While these effects appear to be quite small, it is important to note that a one unit increase in SAT score is actually a 10 point increase. Thus, these coefficients are actually 10 times larger than they appear. As a result, the effect of a one unit increase in SAT Math may actually lead to a smaller improvement in achievement relative to their class of 0.00270 points (the effect of SAT Verbal is a reduction by 0.00174 points). These impacts are still quite small and they are not statistically significant, suggesting

that SAT scores have little to no impact on a student's change in performance in post-requisite college courses.

The instructor characteristics in our model include their gender, race, academic rank, and class size. Among these factors, race, gender, class size, and GTF status are negatively correlated with future student achievement. Of these four characteristics, three exhibit a relationship that is statistically significant. At the 90% confidence level, the coefficient on gender is statistically significant with a magnitude of -0.0557. This implies that having a male instructor can lead to improvement in future student achievement relative to the class that is 0.0557 points lower than if the instructor were female. At the 99% confidence level, race and class size are statistically significant. For race, the coefficient is -0.231. This result suggests that having a non-white instructor may lead to an improvement in future student achievement relative to the class that is 0.231 points lower than if the instructor were white. For class size, the coefficient is -0.00122, implying that increasing the size of a class by one student can decrease the improvement in future student achievement relative to the class by 0.00122 points (an increase of 100 students would suggest a decrease of 0.122 points). The lone relationship that is negative is between GTF status and future student achievement. The magnitude of this relationship is -0.0444. This coefficient suggests that taking a class with a GTF may lead to improvement relative to the class that is 0.0444 points lower than if the class was taken with a tenure-track faculty member. In contrast to the four characteristics that exhibit a negative relationship with future student achievement, being a non-tenure-track Instructor exhibits a positive relationship with future student achievement. The coefficient on this relationship is 0.0264. This suggests that taking a class taught by a non-tenure-track instructor can lead to improvement relative to the class that is 0.0264 points higher than if a tenure-track faculty member taught the class.

In addition to the various factors included in our core future student achievement regression, we also created variations of this core model that included an interaction variable between student and instructor race. The results of the student race-instructor race interaction suggest that a nonwhite student having a white instructor is negatively correlated with the change in these students' achievement (see table 4). The implication of this result is that nonwhite students learn less from white instructors relative to their learning from nonwhite instructors.

Discussion

Before any true discussion or analysis concerning these results can occur, it is important to note the context surrounding the magnitudes of our reported results. Since the maximum numerical course evaluation score is five, reported coefficients that appear quite small may actually have a relatively large impact. This issue is further compounded by the narrow range in which course evaluation scores tend to fall. Across our data, the average score for each of the seven course evaluation questions falls within the range of 4.171 to 4.300 (see table 2). Not only is this range narrow but also it clearly indicates that distribution is skewed towards higher scores (See figures 1-7). To further complicate matters, when course evaluation scores are made public, they are published with only one decimal place included. Given the bunching at the top of the distribution, an increase in evaluation scores of 0.10 points is a noticeable and significant change. The ultimate consequence of this situation is that results that may seem too small to have a practical impact on the surface may in fact be influential.

In terms of magnitude, the most striking result is the impact of grades on SET scores. Though the effect may seem small without context, the fact that an instructor could theoretically increase the grade average in their class from a C to a B and increase their evaluation scores from, say, a 4.4 to a 4.7 is a legitimate cause for concern. Since, as we have seen previously,

course evaluation scores tend to be skewed towards the high end of the distribution, the distinction between a 4.4 and a 4.7 is quite dramatic. Consider once again the variety of uses of SETs. A tenure or award committee would surely view an instructor with a 4.7 SET score as a much more capable instructor than one with a 4.4 even though they are separated by a mere 0.3 points.

Of course, the reported relationship between grades and SET scores is not necessarily causal. In fact, there are multiple hypotheses concerning the root causes of the connection between grades and SET scores that would render this relationship perfectly innocuous. If the validity hypothesis, student characteristics hypothesis, or some combination of the two dominates the grading leniency hypothesis, then the legitimacy of SETs are not threatened by the relationship between grades and SETs. However, it is all but impossible to determine the true driving force behind the relationship. This uncertainty is the source of concern regarding grade-SET relationship. It is possible that the grade connection is a feature of SETs but it is perhaps just as likely that it is a bug.

Class size is one potential influencer of SETs where, when taken at face value, our results seem to indicate that it has a minimal impact on instructor ratings. However, these coefficients, like the -0.000611 on the question of instructor quality, represent the change in SET score due to the addition of a single student to the class. When you begin to add more and more students to a class, this effect becomes larger and larger. Thus, teaching a large lecture style class may considerably reduce SET scores as opposed to teaching a smaller class. As a result, instructors who teach large classes may be consistently penalized by the course evaluation system as it currently exists. Of course, it is certainly possible that teaching a large course actually negatively impacts teaching quality and these observed differences in SET scores are accurately reflecting

this relationship. It could even be some combination of an effect of a flaw in the evaluation system and the negative impacts of teaching large courses that lead to the negative relationship between class size and SET scores. However, it would be nearly impossible to accurately attribute a portion of the relationship to either factor and, as a result, using SETs to compare instructors who teach different sized classes may systematically disadvantage those instructors who teach larger classes. We also observe a negative relationship between class size and future student achievement. Like the effect on SET scores, the effect of class size on future student achievement appears to be small on the surface (a coefficient of -0.00122). However, the same magnitude effect that influences the relationship with SET scores also exists for the relationship with future student achievement. Ultimately, the relationship between class size and future student achievement reaffirms the commonly held belief that students learn less effectively in large classes.

Given these results, perhaps the most important implication of these relationships is the clear evidence that bigger classes are worse for both students and instructors. Instructors receive worse SET scores, indicating that they are less effective teachers when teaching larger classes. Similarly, students achieve less relative to their class after taking large classes, suggesting that these large classes lead to less effective learning. The fact that large classes have negative impacts on both students and their instructors suggest that administrators should exercise caution when making decisions that may increase average class sizes and perhaps should even work actively to reduce class sizes. These negative impacts also suggest that tenure and award committees should take into account class size when considering candidates.

One of our more surprising findings concerned the influence of race on SET scores. This is an area of the existing SET literature where not much research exists. However, our findings

run in contrast to what literature does indeed exist. We find a consistent (though statistical significance is inconsistent) negative relationship between being a white instructor and SET scores. Additionally, our reported relationship between being a non-white instructor and SET scores is both inconsistent and not statistically significant. One possible explanation for this relationship is that the cultural and institutional barriers to entry in academia make it so only the most exceptional non-white instructors can obtain positions. Thus the University's white instructors may in fact be inferior teachers than their non-white peers. However, this theory contradicts our results concerning the impact of having a non-white instructor on future student achievement. These results indicate that students who receive instruction from non-white instructors perform worse in future courses than do students taught by white instructors. However, this relationship may be explained by an effect other than race.

What may better explain our contradictory results concerning the impact of race on SET scores and future student achievement is the distinct possibility that our methodology excluded both poor and exceptional instructors of all races that would have led to different results. Based on this possibility, caution should be exercised when drawing conclusions from these results and further research with more concrete definitions of instructor race (ideally self-reported race) should be performed in the future.

Regarding the relationship between academic rank and SET scores, our results suggest that non-tenure-track instructors receive higher scores than their tenure-track faculty counterparts while GTFs receive lower scores than tenure-track-faculty. While these relationships are not consistent across all seven course evaluation questions, they do exist for the instructor quality and amount learned questions (two of the three questions that have a positive relationship with future student achievement). Based on these results, it appears that tenure-track faculty are better

instructors than GTFs but worse instructors than non-tenure-track faculty. We find that this pattern also holds when looking at future student achievement. There are two main theories that could explain why non-tenure-track instructors are better teachers than tenure-track faculty. One theory posits that instructors who are on the tenure track, particularly those who have reached the upper ranks, may become worse teachers. This decline in teaching quality is most likely not due to an actual loss of teaching ability but rather to a shift in priorities. More highly ranked faculty members may have a greater desire to focus on research, have additional responsibilities within the department or University, or could simply lose interest in teaching (especially if they no longer need high SET scores to facilitate their promotion). Alternatively, it is possible that non-tenure-track Instructors are in fact associated with better student outcomes than tenure-track faculty. This could be due to a greater focus by non-tenure track Instructors on teaching that allows them to accumulate more teaching experience and dedicate more time to their classes and to improving their teaching effectiveness. This effect could be further compounded by the fact that non-tenure-track Instructors must maintain higher levels of teaching quality in order to keep or renew their appointments.

Another notable relationship we found was the effect of academic rank on response rate. We found a positive and statistically significant relationship between both non-tenure-track Instructor status and GTF status. This implies that students who take courses from tenure-track faculty are less likely to complete evaluations than if they take a course from a non-tenure-track Instructor or GTF. One possible explanation for this phenomenon relates back to the idea that non-tenure-track Instructors and GTFs have more at stake due to SETs. This explanation hypothesizes that students, in their decision of whether or not to complete evaluations, take into account the value that their evaluation could bring to their instructor. This consideration could be

sparked by an in-class announcement by their instructor about the importance of SET scores to their professional advancement, or simply by an accumulation of information over time of how these scores are used within the University. No matter how the recognition process occurs, if a student is aware that these evaluations are used in the decision making process for determining promotions, they might be more inclined to complete evaluations. Thus, when students are taught by an instructor who has reached higher academic ranks, they may perceive a loss in value of their evaluations and lose a key incentive towards actually completing them.

Our primary goal in conducting this research was to test the hypothesis that course evaluations at the University of Oregon are valid measures of teaching quality. Of all of our results, two of them are particularly troubling evidence against this hypothesis. The first of these findings concerns the course evaluation questions themselves. Of the seven course evaluation questions that address instructor and course quality, only one question, the question that asks students to evaluate the overall quality of their instructor, exhibited a positive and statistically significant relationship with future student learning. The implication of this finding is that six of the seven course evaluation questions cannot be valid measures of teaching quality. If these questions do not positively correlate with actual future student learning, controlling for a variety of student and instructor characteristics that may influence student learning, then these questions are measuring something other than teaching quality. A negative relationship between a course evaluation question and future student learning suggests that better course evaluation scores, which should reflect some particular element of instructor quality, leads to students learning less in their future courses. For example, the question regarding the instructor's use of class time exhibits a negative relationship that is statistically significant at the 95% confidence level. This means that students who learn from instructors who use class time better actually exhibit less

improvement in future courses. If we believe that good use of class time is in fact a component of quality teaching, then the idea that a question evaluating the use of class time would be negatively correlated with future student achievement is nonsensical. These results are clear evidence against the validity of course evaluations as a measure of teaching quality.

However, the questions that exhibit positive relationships, the questions concerning instructor quality, communication, and the amount learned in the course, may in fact be valid measures of teaching quality. The relationship between the instructor quality question and future student achievement suggests that students taught by higher quality instructors (as measured by the course evaluation score) do better in subsequent courses. Additionally, the positive (though not statistically significant) relationship between the amount a student learns in a course and their future achievement indicates that as a student learns more in a class, they do better in subsequent courses. Both of these relationships align with commonly accepted principles of the teacher-student relationship. We expect that better teachers will impart more skills and knowledge onto their students, which will allow those students to achieve more relative to their peers. Thus, the positive relationships between the instructor quality question and the amount learned question on UO course evaluations suggest that these questions may in fact be valid measures of teaching quality.

The second finding is the influence of gender on both SET scores and future student achievement. In our examination of the relationship between gender and SET score, we found a consistent, negative, and statistically significant relationship between gender and SET scores. These results imply that, since female instructors receive lower SET scores, they are worse instructors than their male counterparts. While the effect is not large enough to lower an instructor's course evaluation score on its own, it is possible that the impact of being female

could have a tipping point effect in that, if an instructor is on the precipice of moving down a score, the effect of being female could push them over the edge. For example, if an instructor, absent the effect of being female, had an evaluation score of 4.26, the effect of being female could bump their score down to the point where the published rounded score was a 4.2 rather than a 4.3. This tipping point effect represents a disadvantage to female instructors due only to the system of evaluation rather than any fault of their own.

While the tipping point effect of gender on SET scores is a cause for concern in and of itself, this concern only grows when looking at the tipping point effect in concert with the effect of being female on future student achievement. We find that having a female instructor in a prerequisite class has a positive effect on student achievement in a post requisite class. This finding indicates that students learn more from female instructors and implies that female instructors are higher quality teachers than their male counterparts. Now consider this relationship with our previously reported relationship between gender and SET scores: female instructors are higher quality instructors yet they receive consistently lower course evaluation scores. This finding, that not only do female instructors receive lower SET scores but they do so in spite of being higher quality instructors, is clear and damning evidence against the validity of course evaluations at the UO. This evidence suggests that course evaluations are biased against female instructors. We cannot lose sight of the ramifications of these findings. SET scores are a critical component of the decision making process for promotion, tenure, merit raises, and teaching awards and yet they systematically disadvantage female instructors. These results, combined with other documented disadvantages to women, paint a bleak picture for women in higher education.

Conclusion

In this paper we have addressed the impact of a variety of factors on SET scores and have explored the validity of SETs as a measure of teaching quality. When looking at the impact of these factors on SET scores, it is not immediately evident that SET scores are affected to the point that they do not function as a valid measure of teaching quality. Some factors hypothesized as potential sources of bias to SET scores, such as instructor race and class level exhibited generally inconclusive results. Others, such as academic rank, class size, and instructor gender, exhibited results that were consistent, though small in magnitude. The effect of grades on SET scores was consistently positive and large in magnitude but the root cause of this effect is difficult to discern. These results undoubtedly raise concerns about the validity of SETs, but they do not themselves provide enough evidence to suggest that SETs are not valid measures of teaching effectiveness.

However, when these results are evaluated together with the results of our investigation into the relationship between SETs and future student achievement, this conclusion changes drastically. Some of our results regarding future student learning are troubling but are not necessarily an indictment of SETs. For example, we find a negative relationship between class size and future student achievement. While this is not necessarily a flaw in SETs, there still exists the possibility that it is. At the very least, it is concerning for those invested in student learning at the UO. We also conclude that non-tenure-track Instructors are better teachers than GTFs and tenure-track faculty.

But it is our final two findings that provide the most compelling case against the validity of SETs. For one, we find that only three of the seven UO course evaluation questions are positively correlated with future student achievement. In other words, the instructor quality

question, communication question, and amount learned question are the only UO course evaluation questions that can be valid measures of teaching quality. To make matters worse, our results suggest that female instructors receive systematically lower course evaluation scores while their students achieve more than their peers taught by male instructors in future courses. This finding is distinct evidence that SETs are biased against female instructors. When combined with the grade effect, class size effect, and the invalidity of the majority of course evaluation questions, it is abundantly clear that course evaluations are not a valid measure of teaching quality at the University of Oregon.

Appendix

Figure 1: Distribution of Question 1 Scores

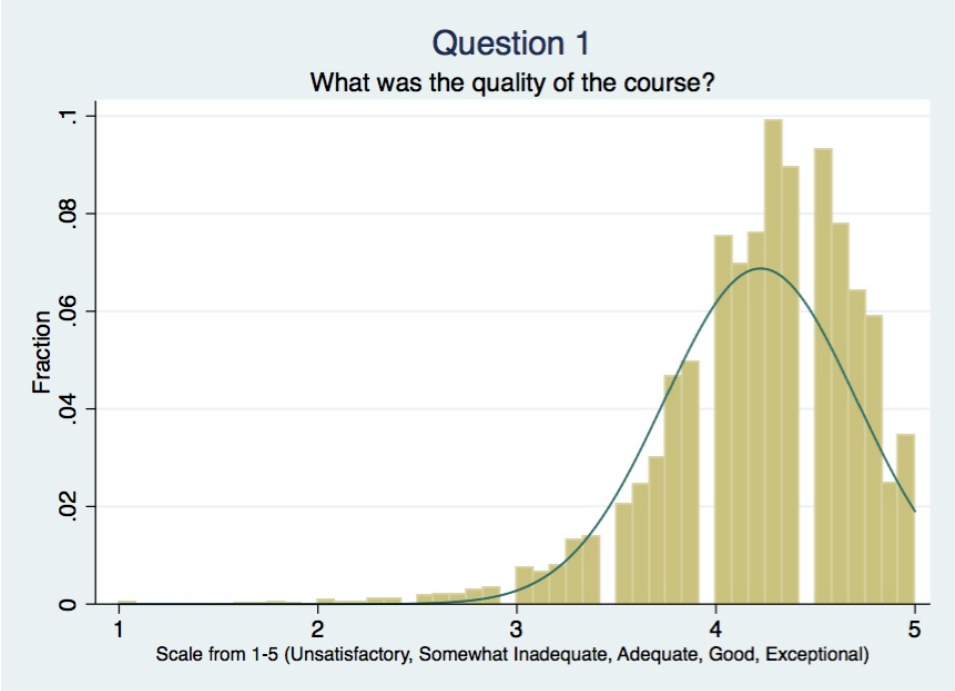

Figure 2: Distribution of Question 2 Scores

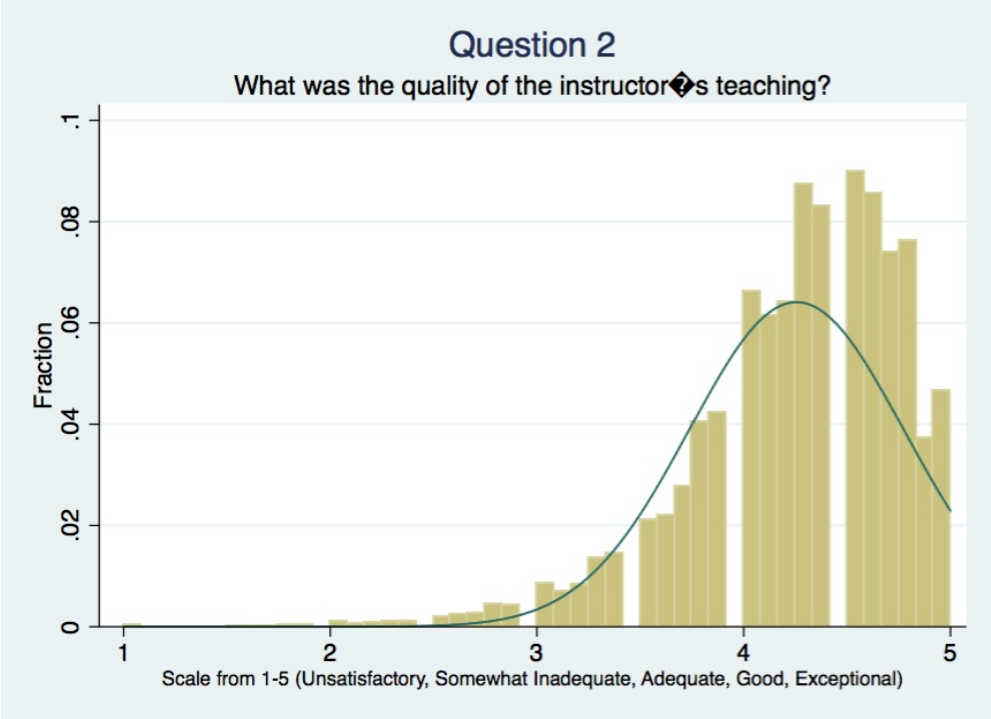

Figure 3: Distribution of Question 3 Scores

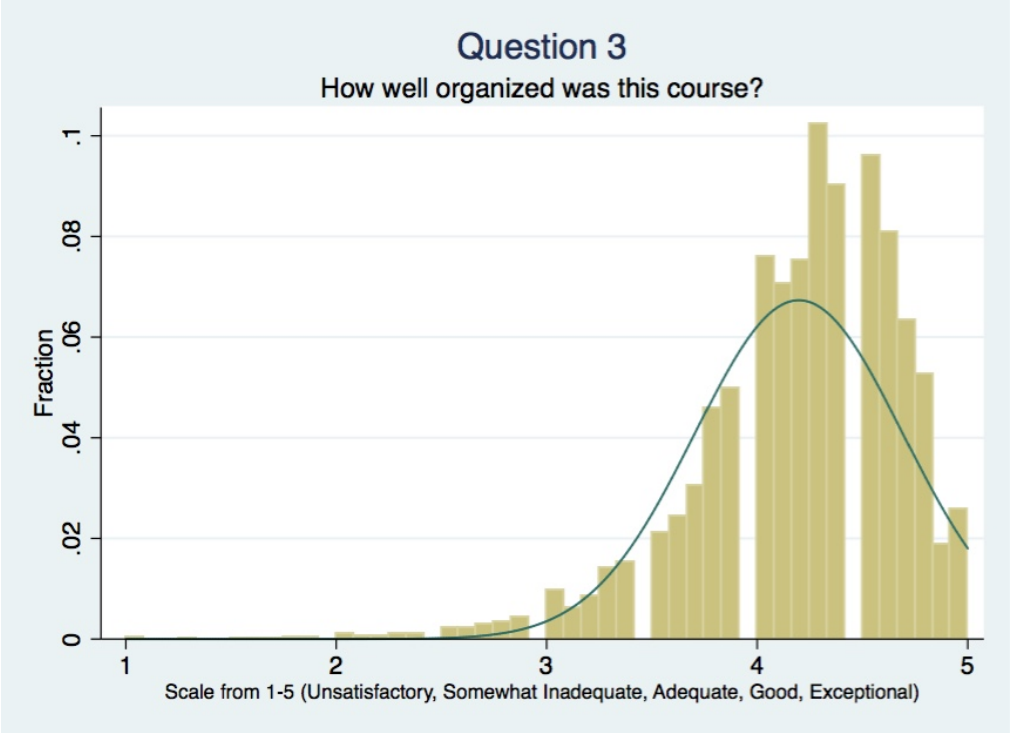

Figure 4: Distribution of Question 4 Scores

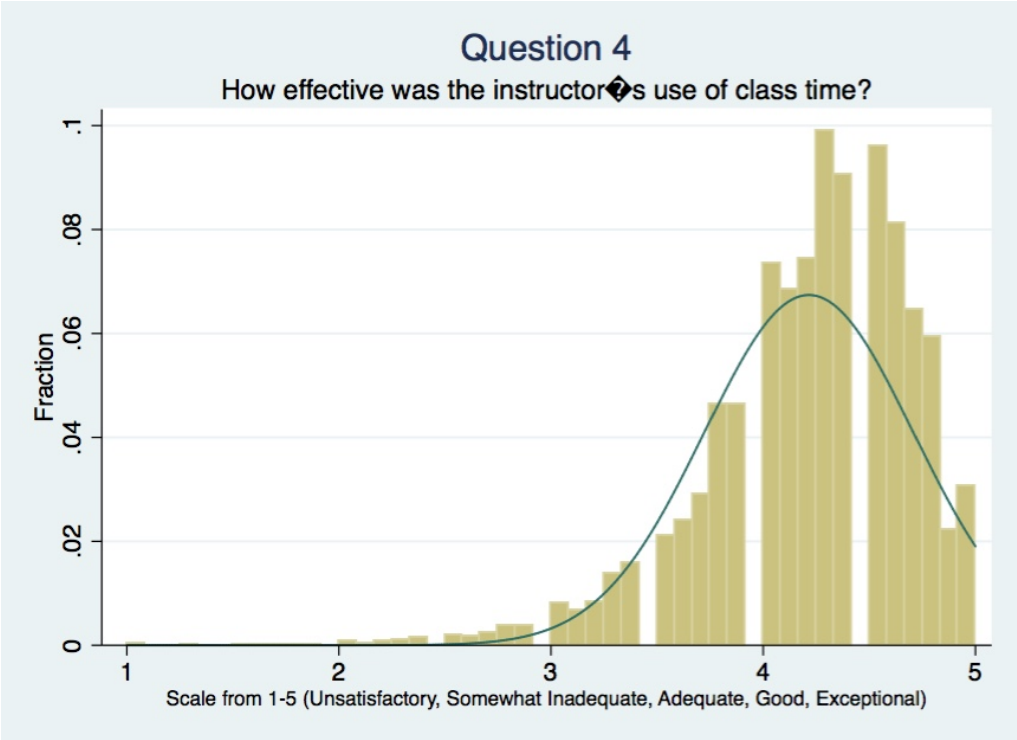

Figure 5: Distribution of Question 5 Scores

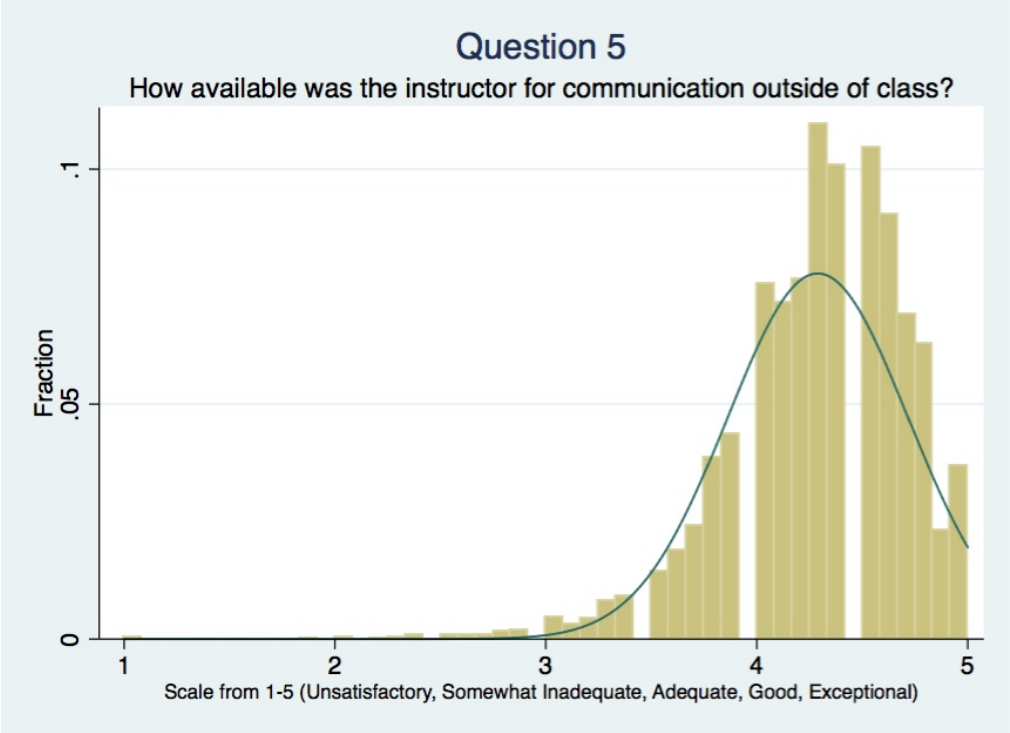

Figure 6: Distribution of Question 6 Scores

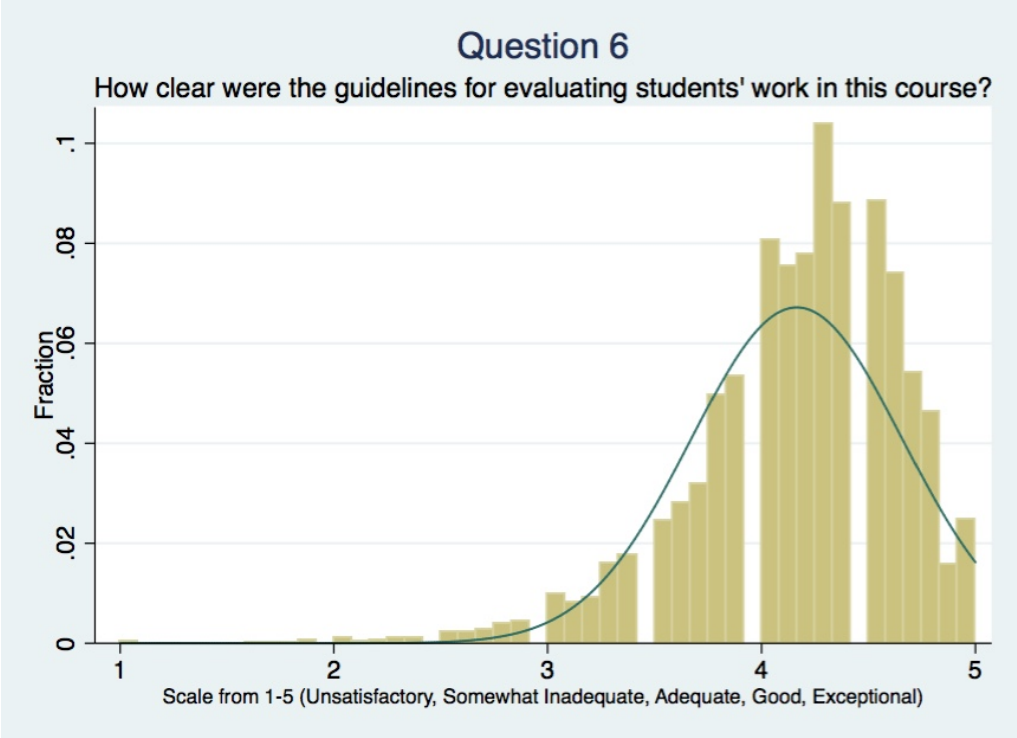

Figure 7: Distribution of Question 7 Scores

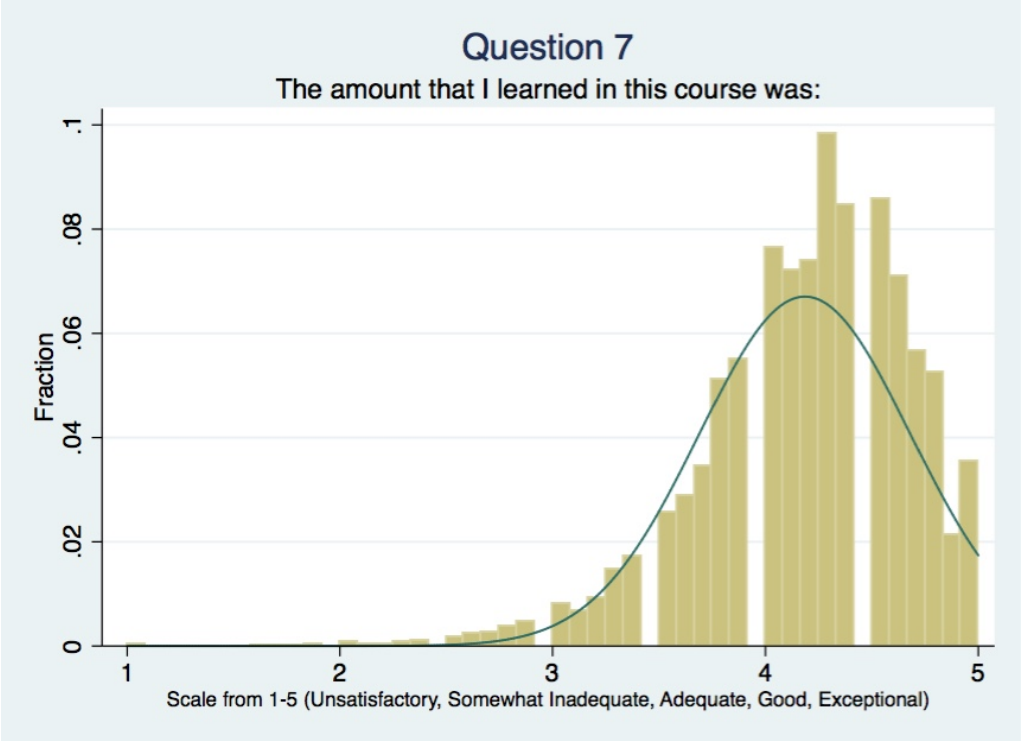

Table 1: Model 1 Results

Column1	(1)	(2)	(3)	(4)	(5)
VARIABLES	Q1: Course Quality	Q2: Instructor Quality	Q3: Course Organization	Q4: Instructor's Usage of Class Time	Q5: Communication
AVGGPA	0.306***	0.319***	0.182***	0.206***	0.240***
	(0.00906)	(0.0103)	(0.00998)	(0.00977)	(0.00825)
INSTRUCTOR GENDER	-0.0498***	-0.0578***	-0.0199***	-0.0266***	-0.0158***
	(0.00573)	(0.00650)	(0.00631)	(0.00618)	(0.00522)
COURSE LEVEL	0.0255***	0.0265***	-0.00623	-0.00617	0.0401***
	(0.00705)	(0.00800)	(0.00776)	(0.00761)	(0.00642)
INSTRUCTOR STATUS	0.0192***	0.0175**	0.0247***	0.0263***	0.0284***
	(0.00735)	(0.00834)	(0.00810)	(0.00793)	(0.00669)
GTF STATUS	-0.0246***	-0.0288***	-0.0127	-0.00292	0.0586***
	(0.00769)	(0.00872)	(0.00847)	(0.00829)	(0.00700)
CLASS SIZE	-0.000643***	-0.000611***	-0.000248***	-0.000420***	-0.000784***
	(5.63e-05)	(6.38e-05)	(6.19e-05)	(6.07e-05)	(5.12e-05)
INSTRUCTOR WHITE	-0.0119**	-0.0109*	-0.0162***	-0.0163***	-0.00300
	(0.00546)	(0.00619)	(0.00601)	(0.00589)	(0.00497)
INSTRUCTOR NON WHITE	-0.00191	-0.0258	0.0361*	0.00723	0.0108
	(0.0170)	(0.0193)	(0.0187)	(0.0183)	(0.0155)
INSTRUCTOR MALE X COURSE AVG GPA					
INSTRUCTOR FEMALE X COURSE AVG GPA					
INSTRUCTOR NONWHITE X COURSE AVG GPA					
INSTRUCTOR WHITE X COURSE AVG GPA					
Constant	3.267***	3.262***	3.627***	3.566***	3.502***
	(0.0302)	(0.0343)	(0.0333)	(0.0326)	(0.0275)
Observations	25,515	25,515	25,515	25,515	25,515
R-squared	0.258	0.241	0.182	0.193	0.235

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 1 continued

Column1	(6)	(7)	(8)	(9)	(10)
VARIABLES	Q6: Clarity of Guidelines	Q7: Amount Learned	Response Rate	Q1 with Interaction	Q2 with Interaction
COURSE AVG GPA	0.307***	0.297***	3.586***	0.275***	0.287***
	(0.00967)	(0.00912)	(0.262)	(0.0132)	(0.0150)
INSTRUCTOR GENDER	-0.0253***	-0.0511***	-0.0151	-0.187***	-0.224***
	(0.00612)	(0.00577)	(0.166)	(0.0495)	(0.0561)
CLASS LEVEL	-0.0359***	0.0346***	-4.526***	0.0259***	0.0267***
	(0.00752)	(0.00709)	(0.204)	(0.00705)	(0.00800)
INSTRUCTOR STATUS	0.0604***	0.0122	0.494**	0.0193***	0.0177**
	(0.00785)	(0.00740)	(0.212)	(0.00736)	(0.00834)
GTF STATUS	0.0562***	-0.0241***	0.824***	-0.0251***	-0.0294***
	(0.00820)	(0.00774)	(0.222)	(0.00769)	(0.00872)
CLASS SIZE	-9.38e-05	-0.000663***	0.00472***	-0.000650***	-0.000619***
	(6.00e-05)	(5.66e-05)	(0.00162)	(5.63e-05)	(6.39e-05)
INSTRUCTOR WHITE	-0.00187	-0.0126**	-0.212	-0.108**	-0.0780
	(0.00582)	(0.00549)	(0.158)	(0.0482)	(0.0546)
INSTRUCTOR NONWHITE	0.0159	0.000109	-0.405	-0.00552	-0.0299
	(0.0181)	(0.0171)	(0.491)	(0.0170)	(0.0193)
INSTRUCTOR MALE X COURSE AVG GPA				0	0
				(0)	(0)
INSTRUCTOR FEMALE X COURSE AVG GPA				0.0428***	0.0518***
				(0.0153)	(0.0174)
INSTRUCTOR NONWHITE X COURSE AVG GPA				0	0
				(0)	(0)
INSTRUCTOR WHITE X COURSE AVG GPA				0.0299**	0.0209
				(0.0150)	(0.0170)
Constant	3.144***	3.271***	53.44***	3.368***	3.364***
	(0.0323)	(0.0304)	(0.873)	(0.0431)	(0.0488)

Observations	25,515	25,515	25,515	25,515	25,515
R-squared	0.224	0.258	0.502	0.258	0.242

Table 2: Model 1 Summary Statistics

Column1	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
VARIABLES	N	Mean	SD	Min	Max	sum_w	Var	Skewness	Kurtosis	Sum
COURSE AVG GPA	27,161	3.209	0.362	1.644	4.300	27,161	0.131	-0.0295	2.753	87,166
INSTRUCTOR GENDER	65,379	0.464	0.499	0	1	65,379	0.249	0.146	1.021	30,315
COURSE LEVEL	70,446	0.454	0.498	0	1	70,446	0.248	0.184	1.034	32,004
INSTRUCTOR STATUS	70,446	0.271	0.445	0	1	70,446	0.198	1.028	2.057	19,122
GTF STATUS	70,446	0.546	0.498	0	1	70,446	0.248	-0.185	1.034	38,467
CLASS SIZE	70,446	31.45	39.72	5	513	70,446	1,577	5.366	42.78	2.216e+06
INSTRUCTOR WHITE	70,446	0.394	0.489	0	1	70,446	0.239	0.434	1.189	27,747
INSTRUCTOR NONWHITE	70,446	0.0508	0.220	0	1	70,446	0.0482	4.092	17.74	3,578
COURSE QUALITY	70,445	4.233	0.473	1	5	70,445	0.223	-0.977	4.829	298,178
INSTRUCTOR QUALITY	70,443	4.264	0.508	1	5	70,443	0.258	-1.098	4.865	300,347
AMOUNT LEARNED	70,437	4.193	0.488	1	5	70,437	0.238	-0.884	4.526	295,316
COURSE ORGANIZATION	70,444	4.200	0.486	1	5	70,444	0.236	-1.128	5.339	295,877
USE OF CLASS TIME	70,440	4.221	0.484	1	5	70,440	0.234	-1.084	5.172	297,294
COMMUNICATION	70,441	4.300	0.414	1	5	70,441	0.172	-0.936	5.226	302,874
COURSE GUIDELINES	70,439	4.171	0.485	1	5	70,439	0.235	-1.006	4.909	293,777

Table 2 continued

Column1	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
VARIABLES	p1	p5	p10	p25	p50	p75	p90	p95	p99
COURSE AVG GPA	2.373	2.627	2.751	2.950	3.201	3.470	3.688	3.810	4
INSTRUCTOR GENDER	0	0	0	0	0	1	1	1	1
COURSE LEVEL	0	0	0	0	0	1	1	1	1
INSTRUCTOR STATUS	0	0	0	0	0	1	1	1	1
GTF STATUS	0	0	0	0	1	1	1	1	1
CLASS SIZE	5	8	10	15	22	30	56	92	219
INSTRUCTOR WHITE	0	0	0	0	0	1	1	1	1
INSTRUCTOR NONWHITE	0	0	0	0	0	0	0	1	1
COURSE QUALITY	2.800	3.400	3.600	4	4.300	4.600	4.800	4.900	5
INSTRUCTOR QUALITY	2.700	3.300	3.600	4	4.300	4.600	4.800	4.900	5
AMOUNT LEARNED	2.700	3.300	3.600	3.900	4.300	4.500	4.800	4.900	5
COURSE ORGANIZATION	2.700	3.300	3.600	4	4.300	4.500	4.700	4.800	5
USE OF CLASS TIME	2.700	3.300	3.600	4	4.300	4.600	4.800	4.900	5
COMMUNICATION	3	3.600	3.800	4.100	4.300	4.600	4.800	4.900	5
COURSE GUIDELINES	2.700	3.300	3.500	3.900	4.200	4.500	4.700	4.800	5

Table 3: Model 1 Key

Variables	Description
COURSE AVG GPA	Average GPA awarded in the course
INSTRUCTOR GENDER	=1 if first name is coded as female; 0 otherwise
COURSE LEVEL	=1 if the course is an upper division course; 0 otherwise
INSTRUCTOR STATUS	=1 if the rank of the educator is an Instructor; 0 otherwise
GTF STATUS	=1 if the educator is unranked (i.e. Graduate Teaching Fellow); 0 otherwise
CLASS SIZE	Number of students enrolled in the class
INSTRUCTOR WHITE	=1 if the probability of the educator being white > .9; 0 otherwise
INSTRUCTOR NONWHITE	=1 if the probability of the educator being a non white race >.9; 0 otherwise
COURSE QUALITY	Dependent Variable. Average evaluation score for Question 1 (What was the Quality of the Course?) on a scale from 1-5 (Unsatisfactory, Somewhat Inadequate, Adequate, Good, Exceptional)
INSTRUCTOR QUALITY	Dependent Variable. Average evaluation score for Question 2 (What was the Quality of the Instructor?) on a scale from 1-5 (Unsatisfactory, Somewhat Inadequate, Adequate, Good, Exceptional)
COURSE ORGANIZATION	Dependent Variable. Average evaluation score for Question 3 (How well organized was this course?) on a scale from 1-5 (Unsatisfactory, Somewhat Inadequate, Adequate, Good, Exceptional)
USE OF CLASS TIME	Dependent Variable. Average evaluation score for Question 4 (How effective was the instructor's use of class time?) on a scale from 1-5 (Unsatisfactory, Somewhat Inadequate, Adequate, Good, Exceptional
COMMUNICATION	Dependent Variable. Average evaluation score for Question 5 (How available was the instructor for communication outside of class?) on a scale from 1-5 (Unsatisfactory, Somewhat Inadequate, Adequate, Good, Exceptional)
COURSE GUIDELINES	Dependent Variable. Average evaluation score for Question 6 (How clear were the guidelines for evaluating students' work in this course?) on a scale from 1-5 (Unsatisfactory, Somewhat Inadequate, Adequate, Good, Exceptional)
AMOUNT LEARNED	Dependent Variable. Average evaluation score for Question 7 (The amount that I learned in this course was:) on a scale from 1-5 (Unsatisfactory, Somewhat Inadequate, Adequate, Good, Exceptional)

Table 4: Model 2 Results

Column1	(1)	(2)	(5)
VARIABLES	Core Model	Core Model with Q2 and Q7	Core Model w/ Interaction
STUDENT GENDER	0.0234	0.0228	0.0223
	(0.0228)	(0.0228)	(0.0228)
STUDENT RACE	-0.0306	-0.0291	-0.0786***
	(0.0241)	(0.0241)	(0.0285)
OUTSTATE	0.0565**	0.0557**	0.0565**
	(0.0222)	(0.0222)	(0.0222)
COLLEGE GPA	0.0251	0.0248	0.0266
	(0.0228)	(0.0228)	(0.0228)
SATM	-0.000270	-0.000267	-0.000269
	(0.000179)	(0.000179)	(0.000179)
SATV	-0.000174	-0.000165	-0.000174
	(0.000153)	(0.000153)	(0.000153)
STUDENT AGE	-0.0111	-0.0110	-0.0107
	(0.0116)	(0.0116)	(0.0116)
INTL	-0.0418	-0.0386	-0.0432
	(0.0773)	(0.0773)	(0.0773)
INSTRUCTOR GENDER	-0.0557*	-0.0442	-0.0553*
	(0.0318)	(0.0308)	(0.0318)
INSTRUCTOR WHITE	0.0464	0.0441	0.0962***
	(0.0294)	(0.0293)	(0.0334)
INTSTRUCTOR NONWHITE	-0.231***	-0.218***	-0.228***
	(0.0602)	(0.0601)	(0.0602)
INSTRUCTOR STATUS	0.0264	0.0283	0.0242
	(0.0420)	(0.0403)	(0.0420)
GTF STATUS	-0.0444	-0.0383	-0.0480
	(0.0468)	(0.0440)	(0.0468)
CLASS SIZE	-0.00122***	-0.00119***	-0.00123***
	(0.000168)	(0.000164)	(0.000168)
COURSE QUALITY	-0.267*	-0.401***	-0.269*
	(0.145)	(0.134)	(0.145)
INSTRUCTOR QUALITY	0.334***	0.244***	0.339***
	(0.101)	(0.0943)	(0.101)

COURSE ORGANIZATION	-0.0566		-0.0568
	(0.0861)		(0.0861)
USE OF CLASS TIME	-0.190**		-0.193**
	(0.0886)		(0.0886)
COMMUNICATION	0.00591		0.00273
	(0.0717)		(0.0717)
COURSE GUIDELINES	-0.0821		-0.0817
	(0.0872)		(0.0872)
AMOUNT LEARNED	0.0291	-0.0144	0.0324
	(0.0973)	(0.0962)	(0.0972)
STUDENT NONWHITE X INSTRUCTOR WHITE			0
			(0)
STUDENT NONWHITE X INSTRUCTOR WHITE			-0.157***
			(0.0501)
STUDENT WHITE X INSTRUCTOR NONWHITE			0
			(0)
STUDENT WHITE X INSTRUCTOR WHITE			0
			(0)
Constant	1.029**	0.752**	1.050***
	(0.325)	(0.296)	(0.324)
Observations	10,094	10,094	10,094
R-squared	0.124	0.123	0.125
Standard errors in parentheses			
*** p<0.01, ** p<0.05, * p<0.1			

Table 5: Model 2 Summary Statistics

Column1	Column2	Column3	Column4	Column5	Column6	Column7	Column8	Column9	Column10	Column11
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
VARIABLES	N	Mean	SD	Min	Max	sum_w	Var	Skewness	Kurtosis	Sum
DIFFERENCE	1,554	-0.755	1.416	-5.130	4.055	1,554	2.004	-0.319	3.570	-1,173
STUDENT GENDER	1,554	0.451	0.498	0	1	1,554	0.248	0.197	1.039	701
STUDENT RACE	1,554	0.600	0.490	0	1	1,554	0.240	-0.410	1.168	933
OUTSTATE	1,554	0.569	0.495	0	1	1,554	0.245	-0.281	1.079	885
COLLEGE GPA	1,554	3.024	0.492	0.900	4.300	1,554	0.242	-0.362	3.374	4,699
SATM	1,218	560.5	80.89	320	800	1,218	6,544	-0.0337	2.774	682,690
SATV	1,218	545.1	89.18	310	800	1,218	7,953	0.0780	3.080	663,880
STUDENT AGE	1,554	18.94	1.779	17	39	1,554	3.164	4.868	39.02	29,432
INTL	1,554	0.182	0.386	0	1	1,554	0.149	1.647	3.714	283
INSTRUCTOR GENDER	1,415	0.805	0.396	0	1	1,415	0.157	-1.539	3.369	1,139
INSTRUCTOR WHITE	1,554	0.0219	0.146	0	1	1,554	0.0214	6.537	43.73	34
INSTRUCTOR NONWHITE	1,554	0.00450	0.0670	0	1	1,554	0.00449	14.80	220.0	7
INSTRUCTOR STATUS	1,554	0	0	0	0	1,554	0			0
GTF STATUS	1,554	0	0	0	0	1,554	0			0
CLASS SIZE	1,554	310.8	116.9	18	441	1,554	13,654	-0.749	2.389	482,993
COURSE QUALITY	1,554	3.626	0.592	2.800	4.900	1,554	0.351	-0.106	1.556	5,635
INSTRUCTOR QUALITY	1,554	3.593	0.712	2.600	5	1,554	0.506	-0.102	1.551	5,584
COURSE ORGANIZATION	1,554	3.836	0.511	3.200	4.800	1,554	0.261	0.113	1.496	5,961
USE OF CLASS TIME	1,554	3.831	0.482	3.200	4.900	1,554	0.233	0.147	1.676	5,953
COMMUNICATION	1,554	3.732	0.456	3.200	4.700	1,554	0.208	0.420	1.805	5,800
COURSE GUIDELINES	1,554	3.697	0.535	3	4.800	1,554	0.286	0.0860	1.513	5,744
AMOUNT LEARNED	1,554	3.664	0.508	3	4.800	1,554	0.258	0.0559	1.520	5,693

Table 5 continued

Column1	Column12	Column13	Column14	Column15	Column16	Column17	Column18	Column19	Column20
	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
VARIABLES	p1	p5	p10	p25	p50	p75	p90	p95	p99
DIFFERENCE	-4.430	-3.430	-2.730	-1.446	-0.564	0.128	0.763	1.422	2.748
STUDENT GENDER	0	0	0	0	0	1	1	1	1
STUDENT RACE	0	0	0	0	1	1	1	1	1
OUTSTATE	0	0	0	0	1	1	1	1	1
COLLEGE GPA	1.730	2.200	2.390	2.720	3.040	3.350	3.660	3.800	4
SATM	380	420	455	510	555	620	660	690	750
SATV	340	400	430	490	550	600	660	690	770
STUDENT AGE	18	18	18	18	18	19	20	22	27
INTL	0	0	0	0	0	0	1	1	1
INSTRUCTOR GENDER	0	0	0	1	1	1	1	1	1
INSTRUCTOR WHITE	0	0	0	0	0	0	0	0	1
INSTRUCTOR NONWHITE	0	0	0	0	0	0	0	0	0
INSTRUCTOR STATUS	0	0	0	0	0	0	0	0	0
GTF STATUS	0	0	0	0	0	0	0	0	0
CLASS SIZE	33	100	139	268	332	389	441	441	441
COURSE QUALITY	2.800	2.800	2.800	3.200	3.700	4.200	4.300	4.400	4.600
INSTRUCTOR QUALITY	2.600	2.600	2.600	3.100	3.500	4.200	4.400	4.500	4.800
COURSE ORGANIZATION	3.200	3.200	3.200	3.400	3.700	4.300	4.500	4.600	4.800
USE OF CLASS TIME	3.200	3.200	3.200	3.500	3.700	4.300	4.500	4.500	4.800
COMMUNICATION	3.200	3.200	3.200	3.400	3.700	4.100	4.500	4.500	4.600
COURSE GUIDELINES	3	3	3	3.300	3.600	4.200	4.400	4.500	4.500
AMOUNT LEARNED	3	3	3	3.300	3.700	4.100	4.300	4.400	4.500

Table 6: Model 2 Key

Variables	Definition
DIFFERENCE	Dependent Variable. A measure of the change in student achievement from a prerequisite course to a postrequisite course, normalized for the grade distribution of each class
STUDENT GENDER	=1 if the student is female; 0 otherwise
STUDENT RACE	=1 if the student is white; 0 otherswise
OUTSTATE	=1 if the student is a non resident; 0 otherwise
COLLEGE GPA	Student's cumulative college GPA
SATM	Student's Math SAT score or SAT equivalent score
SATV	Student's Verbal SAT score or SAT equivalent score
STUDENT AGE	Age of the student
INTL	=1 if the student is an international student; 0 otherwise
INSTRUCTOR GENDER	=1 if the educator is male; 0 otherwise
INSTRUCTOR WHITE	=1 if the probability of the educator being white > .9; 0 otherwise
INSTRUCTOR NONWHITE	=1 if the probability of the educator being a non white race >.9; 0 otherwise
INSTRUCTOR STATUS	=1 if the rank of the educator is an Instructor; 0 otherwise
GTF STATUS	=1 if the educator is unranked (i.e. Graduate Teaching Fellow); 0 otherwise
CLASS SIZE	Number of students enrolled in the class
COURSE QUALITY	Average evaluation score for Question 1 (What was the Quality of the Course?) on a scale from 1-5 (Unsatisfactory, Somewhat Inadequate, Adequate, Good, Exceptional)
INSTRUCTOR QUALITY	Average evaluation score for Question 2 (What was the Quality of the Instructor?) on a scale from 1-5 (Unsatisfactory, Somewhat Inadequate, Adequate, Good, Exceptional)
COURSE ORGANIZATION	Average evaluation score for Question 3 (How well organized was this course?) on a scale from 1-5 (Unsatisfactory, Somewhat Inadequate, Adequate, Good, Exceptional)
USE OF CLASS TIME	Average evaluation score for Question 4 (How effective was the instructor’s use of class time?) on a scale from 1-5 (Unsatisfactory, Somewhat Inadequate, Adequate, Good, Exceptional)
COMMUNICATION	Average evaluation score for Question 5 (How available was the instructor for communication outside of class?) on a scale from 1-5 (Unsatisfactory, Somewhat Inadequate, Adequate, Good, Exceptional)
COURSE GUIDELINES	Average evaluation score for Question 6 (How clear were the guidelines for evaluating students' work in this course?) on a scale from 1-5 (Unsatisfactory, Somewhat Inadequate, Adequate, Good, Exceptional)
AMOUNT LEARNED	Average evaluation score for Question 7 (The amount that I learned in this course was:) on a scale from 1-5 (Unsatisfactory, Somewhat Inadequate, Adequate, Good, Exceptional)

Table 7: Race Frequencies

Race	Frequency	Percentage
WHITE	27,745	39.38
BLACK	83	0.12
HISPANIC	661	0.94
ASIAN	2,834	4.02
UNCLASSIFIED	39,123	55.54

Works Cited

ACT-SAT Concordance Tables. (2009). ACT Research & Policy. From: http://www.act.org/content/dam/act/unsecured/documents/ACTCollegeBoardJointStatement.pdf

Aleamoni, L. M. (1987). Typical faculty concerns about student evaluation of teaching. *New Directions for Teaching and Learning, 1987*(31), 25-31.

Aleamoni, L. M., & Graham, M. H. (1974). The relationship between CEQ ratings and instructor's rank, class size, and course level. *Journal of Educational Measurement, 11*(3), 189-202.

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), 233-235.

Aleamoni, L. M., & Yimer, M. (1973). An investigation of the relationship between colleague rating, student rating, research productivity, and academic rank in rating instructional effectiveness. *Journal of Educational Psychology, 64*(3), 274.

Anderson, K. J., & Smith, G. (2005). Students' preconceptions of professors: Benefits and barriers according to ethnicity and gender. *Hispanic Journal of Behavioral Sciences, 27*(2), 184-201.

Beleche, T., Fairris, D., & Marks, M. (2012). Do course evaluations truly reflect student learning? Evidence from an objectively graded post-test. *Economics of Education Review, 31*(5), 709-719.

Beran, T., & Violato, C. (2005). Ratings of university teacher instruction: How much do student and course characteristics really matter? *Assessment & Evaluation in Higher Education, 30*(6), 593-601.

Boring, A., Ottoboni, K., & Stark, P. B. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*.

Brandenburg, D. C., & Aleamoni, L. M. (1976). Illinois Course Evaluation Questionnaire (CEQ): Results interpretation manual, Form 73. *Urbana, Ill.: Measurement and Research Division, Office of Instructional Resources, University of Illinois*.

Brandenburg, D. C., Slinde, J. A., & Batista, E. E. (1977). Student ratings of instruction: Validity and normative interpretations. *Research in Higher Education, 7*(1), 67-78.

Centra, J. A. (1981). Research productivity and teaching effectiveness. *ETS Research Report Series, 1981*(1).

Centra, J. A., & Creech, F. R. (1976). The relationship between student, teacher, and course characteristics and student ratings of teacher effectiveness. *Project report, 761*.

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

Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature. *Journal of Marketing Education, 31*(1), 16-30.

Delaney Jr, E. L. (1976). The Relationships of Student Ratings of Instruction to Student, Instructor and Course Characteristics.

Elmore, P. B., & LaPointe, K. A. (1974). Effects of teacher sex and student sex on the evaluation of college instructors. *Journal of Educational Psychology, 66*(3), 386.

Feldman, K. A. (1978). Course characteristics and college students' ratings of their teachers: What we know and what we don't. *Research in Higher Education, 9*(3), 199-242.

Feldman, K. A. (1983). Seniority and experience of college teachers as related to evaluations they receive from students. *Research in Higher Education, 18*(1), 3-124.

Feldman, K. A. (2007). Identifying Exemplary Teachers and Teaching: Evidence from Student Ratings1 *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 93-143): Springer.

Figlio, D. N., Schapiro, M. O., & Soter, K. B. (2015). Are tenure track professors better teachers?. *Review of Economics and Statistics*, *97*(4), 715-724.

Greenwald, A. G., & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American psychologist, 52*(11), 1209.

Hamermesh, D. S., & Parker, A. (2005). Beauty in the classroom: Instructors' pulchritude and putative pedagogical productivity. *Economics of Education Review, 24*(4), 369-376.

King, A. P. (1971). The self-concept and self-actualization of university faculty in relation to student perceptions of effective teaching.

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.

Linsky, A. S., & Straus, M. A. (1975). Student evaluations, research productivity, and eminence of college faculty. *The Journal of Higher Education, 46*(1), 89-102.

MacNell, L., Driscoll, A., & Hunt, A. N. (2015). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education, 40*(4), 291-303.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International journal of educational research, 11*(3), 253-388.

Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319-383): Springer.

Marsh, H. W., & Dunkin, M. J. (1992). Students' evaluations of university teaching: A multidimensional perspective. *Higher education: Handbook of theory and research, 8*, 143-233.

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.

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

Nevill, D. D., Ware, W. B., & Smith, A. B. (1978). A comparison of student ratings of teaching assistants and faculty members. *American Educational Research Journal, 15*(1), 25-37.

Remmers, H. H., & Brandenburg, G. C. (1927). Experimental data on the Purdue rating scale for instructors. *Educational Administration and Supervision*, *13*(6), 399-406.

Smith, B. P. (2007). Student ratings of teaching effectiveness: An analysis of end-of-course faculty evaluations. *College Student Journal, 41*(4), 788.

Stapleton, R. J., & Murkison, G. (2001). Optimizing the fairness of student evaluations: A study of correlations between instructor excellence, study production, learning production, and expected grades. *Journal of Management Education, 25*(3), 269-291.

Statistical Profile of Persons Receiving Doctor's Degrees, By Field of Study and Selected Characteristics: 2012-13 and 2013-14. (2015). from https://nces.ed.gov/programs/digest/d15/tables/dt15_324.80.asp

UO Online Course Evaluations. from https://registrar.uoregon.edu/course-evaluations

Uttl, B., & Smibert, D. (2017). Student evaluations of teaching: teaching quantitative courses can be hazardous to one's career. *PeerJ, 5*, e3299.

Uttl, B., White, C. A., & Gonzalez, D. W. (2016). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*.

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