THE JOURNAL OF ECONOMIC EDUCATION
2018, VOL. 49, NO. 4, 324–339
https://doi.org/10.1080/00220485.2018.1500957

Routledge
Taylor & Francis Group

RESEARCH IN ECONOMIC EDUCATION

Measuring faculty teaching effectiveness using conditional fixed effects

Maia Linask and James Monks

Robins School of Business, University of Richmond, VA, USA

ABSTRACT

Using a dataset of 48 faculty members and 88 courses over 26 semesters, the authors estimate Student Evaluation of Teaching (SET) ratings that are conditional on a multitude of course, faculty, and student attributes. They find that ratings are lower for required courses and those where students report a lower prior level of interest. Controlling for these variables substantially alters the SET ratings for many instructors. The average absolute value of the difference between the faculty ratings controlling just for time effects and fully conditional ratings is nearly one-half of a standard deviation in the students' rating of how much they learned. This difference produces a change in quartile rank for over half the sample across two summary course evaluation measures.

KEYWORDS
assessment; teaching evaluation

JEL CODES
A20; A22

Virtually every college and university in the United States uses student evaluation of teaching (SET) instruments as its primary, if not sole, method of assessing the teaching performance and effectiveness of its faculty (see Baldwin and Blattner [2003], Campbell, Steiner, and Gerdes [2005], and Becker, Bosshardt, and Watts [2012] specifically for economics departments). Most institutions compare average (or median) scores across faculty on a series of SET questions pertaining to various aspects of the course and the faculty member's teaching practices within that course (see Stark and Freishtat [2014] for a critique of using average scores).[1] Indeed, an individual faculty member's teaching effectiveness often is assessed based on the average score on some kind of summary question.[2] The reliability of these scores for identifying effective teaching is important as they are used in tenure and promotion decisions and annual salary assessments, they are closely monitored by university administrators and accreditors, and they impact faculty choices about pedagogy and content.

Our analysis focuses on the limitations of comparing raw SET scores to evaluate teaching effectiveness for personnel decisions, as is done in many institutions. Using data from a single school within a university in the United States, we provide an alternative approach that offers context for individual faculty members' teaching evaluations. To determine the extent of contextual influence prevalent in traditional measures of teaching effectiveness, we compare conditional faculty fixed effects that control for multiple factors outside the instructor's control with baseline faculty fixed effects that control only for semester effects, as is the standard practice in many institutions.[3] We use these faculty fixed effects to calculate the rankings and focus particularly on how students' desire to enroll in a course, as measured by the students' prior interest and whether the course is required or elective, impact rankings quartiles of faculty.[4]

CONTACT James Monks jmonks@richmond.edu Professor of Economics, Robins School of Business, 28 Westhampton Way, University of Richmond, VA 23173, USA.

The analysis produces four main findings. First, we find that students' enthusiasm for taking a course, as measured by self-reported level of interest and whether the course is required or elective, significantly influences student evaluations of teaching quality. Second, we find that including course fixed effects has a substantial impact on the teaching evaluations. This suggests that subject matter and course content, as well as preconceptions and beliefs about courses and even whole disciplines, substantially change the ratings of faculty effectiveness.[5] Third, our results confirm previous findings that raw SET scores are influenced by institutional factors unrelated to an instructor's teaching ability, including class size, class meeting time and frequency, class composition by student cohort and gender, and students' average grade point average (GPA). Finally, we find that, while rankings using SET scores that control only for semester or time effects are positively correlated to rankings that control for all the other factors, there are significant differences in the position of many faculty members in the two orderings.

Using faculty fixed effects conditional on students' interest, course-level fixed effects, and a series of other variables outside the instructor's control changes the ranking of over half of the instructors by enough to move them to a different rankings quartile. For almost a third of the faculty in our sample, the conditional faculty fixed effects move them from below to above the median rating on one summary question, or vice versa. Thus, using raw SET scores alone to evaluate teaching effectiveness without controlling for course and semester attributes overlooks important environmental factors that can provide meaningful context when assessing faculty performance. Furthermore, the importance of course fixed effects implies that producing a single ranking, or even a single mean for comparison, for faculty who teach different courses or teach in different departments is problematic. Rankings and ratings of teaching effectiveness that cross not only disciplines and departments but also schools or divisions can only exacerbate this problem.

We argue that estimating conditional scores that control for as many relevant factors as possible provides useful additional information to using raw scores for making personnel decisions. We do not, however, claim that these conditional faculty fixed effects precisely measure teaching effectiveness. Indeed, the conditional scores that we produce reflect not only the quality of instruction that does not vary over time but also any other instructor attributes that do not vary over time. Furthermore, these conditional fixed effects do not account for changes in teaching effectiveness over time or in different situations. For example, an instructor may be particularly good at teaching an advanced elective course and yet ineffective at teaching introductory courses; even conditional fixed effects do not reflect such distinctions. But given the current importance of SET ratings and rankings in the careers of university and college faculty, using conditional faculty fixed effects in addition to raw scores at least highlights the role that exogenous institutional factors play on the ratings and the resulting comparisons across faculty. Our results suggest that controlling for factors such as students' prior level of interest and whether a course is required is especially important. Sufficient data also would allow evaluating how a faculty member's teaching effectiveness in a given course changes over time and controlling for course fixed effects in calculating ratings and rankings.

While other empirical studies have identified and estimated the influence of many other factors on SET scores, they have not analyzed the degree to which ratings and rankings that are conditional on various observable factors differ from those that rely on the raw SET scores typically used for evaluation of teaching effectiveness. Our study aims to fill that gap by providing evidence of the extent to which SET ratings and rankings change when conditioned on those course and semester characteristics over which the instructor has no influence. The results show that the use of raw SET ratings semester by semester, as is common practice, can lead to significant differences in the evaluation of teaching effectiveness.

Literature review

Studies of student evaluations of teaching have evolved into a large literature that spans multiple disciplines and encompasses both empirical and theoretical papers (see Feldman [2007], Benton

and Cashin [2012], Spooren, Brockx, and Mortelmans [2013], and Clayson [2015] for recent reviews).[6] Within this large literature, our article is most closely related to McPherson (2006), Carrell and West (2010), and Ragan and Walia (2010).

To our knowledge, McPherson (2006) is the only other paper to suggest using estimated faculty fixed effects as an alternative to ranking instructors by their raw SET ratings. Like our article, he uses panel data and instructor and course fixed effects; unlike our article, his primary focus is on the impact of the expected grade, and he includes only five additional controls: number of class meetings per week, share of students who are majors, class size, SET response rates, and instructor experience.[7] In particular, he does not control for whether a course is required or students' prior level of interest, two variables that seem to significantly influence SETs in our data.

In comparing the rankings based on raw scores and fixed effects, McPherson finds that the differences are small for most instructors, with only a few exceptions; the correlation between the unconditional scores and the fixed effect estimates is quite high (0.95 and 0.88 for principles and upper-division courses, respectively). Our results are in stark contrast: we find that controlling for a range of variables, including course fixed effects and students' prior interest in the course, changes the quartile ranking for more than half the faculty in our dataset. Crucially, McPherson uses data from a large university with class sizes up to 300 and analyzes the principles and upper-level courses separately.[8] While this approach does reflect the reality that students often give large, introductory class instructors lower SET ratings, it is not consistent with the common practice of comparing SET ratings for faculty across an entire department or school regardless of the courses taught. Consequently, the finding that rankings within a subset of classes (e.g., large principles classes) are stable when controlling for course and semester variables does not necessarily imply that those rankings are stable across larger sets of classes, which is typically how SET ratings are used for personnel decisions.[9]

Ragan and Walia (2010) consider students' prior interest as well as class size, response rate, the level of a course, and instructor experience in their estimates of the determinants of SET ratings. Like our results, they find that a higher level of prior interest in the course significantly and positively impacts evaluation scores. Ragan and Walia consider principles and nonprinciples courses separately (as does McPherson [2006]) and conclude that ratings using conditional faculty fixed effects are consistent across courses for individual faculty members and that, in general, faculty members who teach principles courses receive higher ratings.[10] They do not, however, provide evidence of how well the raw SET scores that are used for evaluation at most colleges and universities actually reflect the conditional faculty fixed effects that they estimate. Furthermore, as with McPherson, the data in Ragan and Walia are from a larger university (maximum class size is 175) that covers a shorter time span and fewer faculty and courses.

Carrell and West (2010) use random effects and variation in students' grades to estimate the educational "value added" by each instructor. They also compare rankings based on these estimates, which they argue are a proxy for true teaching effectiveness, to raw SET scores. They find that the correlation between the SET score and value-added in the instructor's own class is 0.36, while the correlation between the instructor's SET score and value-added in subsequent related classes is −0.31. However, implementing the approach in Carrell and West requires individual student grades and evaluations and also random assignment of students to classes.[11] Because very few institutional environments meet these stringent requirements, their approach provides little practical guidance for institutions that use SET scores in tenure, promotion, and salary decisions. Finally, because Carrell and West use course-semester random effects, the analysis provides no evidence of the importance of specific course attributes (e.g., whether a course is required or not) or how such course attributes affect rankings. The data also do not include information about the faculty member's preps, course load, or student load, so the impact of these elements is also unknown, and controlling for them could significantly change the random effects estimates.[12]

Many other studies have used both qualitative and quantitative approaches to criticize SET practices on various grounds. For example, Onwuegbuzie et al. (2007) challenge the content validity of SETs because they often omit questions about those qualities that students consider most important for effective teaching, while Marks (2000) challenges the construct validity because students' responses on SETs often confound how much they learn with how much they like a class or instructor.[13] The link between SETs and grade inflation, both theoretical and empirical, is another frequent basis for criticism of the instruments (Langbein 2008). Another concern is the relationship between both student and faculty demographic characteristics, such as gender, race, and ethnicity, and the SET scores. Indeed, Sproule (2002) develops a theoretical model to show that violating any one of three questionable assumptions will lead to a high probability that SET ratings are biased. There is even evidence to suggest that SET scores are correlated with factors such as faculty appearance (Hamermesh and Parker 2005) and weather conditions on the day evaluations are administered (Braga, Paccagnella, and Pellizzari 2014).

A large literature explores the many factors that may bias student evaluations (for some recent examples, see Langbein [2008], McPherson, Jewell, and Kim [2009], Fenn [2015], and Wagner, Rieger, and Voorvelt [2016]). This literature generally groups the potential sources of bias into three categories: student characteristics (e.g., GPA, expected grade), course characteristics (e.g., required or elective, lower- or upper-level, number of students), and instructor characteristics (e.g., gender, race, accent). Previous findings about the relationship between class size and SET ratings are inconclusive, with some papers finding no significant relationship (McPherson, Jewell, and Kim 2009) while others find a positive (although small) effect (De Witte and Rogge 2011) and still others find a negative effect (Isley and Singh 2005). Dilts (1980), Krautmann and Sander (1999), Isley and Singh (2005), Ewing (2012), and Boring, Ottoboni, and Stark (2016) all find that the average expected grade is positively correlated with better evaluations, while De Witte and Rogge (2011) find the opposite. Most papers that analyze gender bias find significant evidence of its existence, including Gramlich and Greenlee (1993), Fenn (2015), Wagner, Rieger, and Voorvelt (2016), Boring, Ottoboni, and Stark (2016) and Boring (2017). While Feldman (1993) finds that the sex of the instructor has no global effect on SETs, this result is quite rare. Drawing on this literature, in our article we control for a thorough set of 16 different course-section and faculty-semester specific variables.[14] We pay particular attention to the effect of students' desire to take a course (as proxied by the level of prior interest and whether the course is required or an elective) on SETs. Those few papers that address this factor, notably Monks and Schmidt (2011), Ragan and Walia (2010), and Olivares (2001), are thus especially relevant to our article.

Finally, in comparing two different proxy measures for teaching effectiveness, our study also relates to papers that propose alternatives to using raw SET scores as a measure of teaching effectiveness. McPherson and Jewell (2007) produce adjusted ratings by setting some of the control variables (e.g., instructor race or average expected course grade) equal to a common value for all instructors to eliminate the bias of those variables on ratings. (See also McPherson [2006] and Ewing [2012], both of which adjust raw SET scores by subtracting the estimated impact, based on regression analysis, of problematic factors.) De Witte and Rogge (2011) propose a new method for calculating measures of teaching effectiveness that entails weighting average scores for an instructor to maximize strengths and minimize weaknesses, resampling to minimize the impact of measurement error, and conditioning on exogenous characteristics.[15]

Weinberg, Hashimoto, and Fleisher (2009) propose a weighted average of student learning and the student's course experience as a better indication of faculty performance in the classroom. Driscoll and Cadden (2010) offer a simpler approach, arguing that reference measures to which SET scores are compared should be department-specific in order to ameliorate the potential bias caused by discipline-specific attributes such as a focus on quantitative skills. Finally, Baldwin and Blattner (2003) recommend supplementing SETs with substantial additional evidence of teaching

effectiveness, such as classroom observation, repeated feedback from students, and a comprehensive teaching portfolio that includes teaching materials.

Data and econometric specification

The data for this study come from administrative records and student course evaluations at a single university in the United States. The faculty and courses utilized in this analysis are restricted to one particular undergraduate school within this university, as the student course evaluation instrument is specific to the school. The sample period covers 26 consecutive semesters, as these are the data made available to us. The sample includes 48 individual faculty members, 88 separate courses, and 1928 course-sections; the panel is unbalanced, as there were some departures and additions to the faculty over the sample period. The unit of observation is the course-section.[16] The dataset preserves the anonymity of each faculty member by numbering each faculty member 1 to 48 and numbering each course 1 to 88.

The SET questions examined in this article are the average course-section ratings for two questions from the student course evaluations: overall instructor rating (1–5, with 5 being the best) and amount learned (1–5, with 5 being the most). Our dependent variables are the arithmetic means of the student ratings for a course-section for each of these questions. Although the SET includes many other questions (see table 1), we restrict our analysis to these two summary questions. A significant share of other institutions also include similar summary questions on their SETs (Becker and Watts 1999; Becker, Bosshardt, and Watts 2012). Furthermore, d'Apollonia and Abrami (1997) argue that even multi-dimensional SETs may "measure a global component" (p.

Table 1. Summary measures.

	Mean	Std. Dev.
Class size	23.39	6.86
Student load	63.46	18.75
Early	0.098	0.297
Late	0.284	0.451
50-minute class	0.312	0.463
Proportion male students	0.601	0.151
Male student * male faculty	0.516	0.256
Proportion sophomores	0.226	0.315
Proportion junior	0.342	0.325
Proportion senior	0.372	0.376
GPA	3.18	0.142
Expected grade	3.05	0.292
Interest in this subject prior to the course	3.47	0.545
Course workload	3.39	0.463
Course level of difficulty	3.66	0.501
Level of critical & analytical thinking	4.00	0.447
Pace of course	3.15	0.207
Number of evals. completed by student	2.66	0.824
Clear and understandable presentation	4.09	0.485
Effectiveness of teaching methods	4.05	0.500
Instructor's daily preparation for class	4.50	0.361
Effectiveness in stimulating interest	3.94	0.514
Enthusiasm for teaching course	4.48	0.366
Availability outside the classroom	4.41	0.370
Respect for students in the class	4.34	0.425
Graded material vs. course content	3.98	0.499
Adequacy of comments on student work	3.76	0.483
Timeliness of feedback on student work	4.26	0.496
Usefulness of text, etc. to learning	3.84	0.512
Amount learned as a result of course	4.12	0.449
Instructor's overall teaching ability	4.18	0.479
Overall quality of course	4.00	0.475

1201). Thus, both because they are often used for evaluating teaching and because there is evidence that they summarize multiple dimensions of teaching, we focus on these two summary questions in our analysis of the use of SETs as indicators of teaching effectiveness.

The explanatory variables fall into several categories.[17] First, we include course-section specific variables that are commonly used in the literature: the (self-reported) average GPA across students in the section, the gender mix of the section, the gender mix interacted with the instructor's sex, the cohort mix of the section, class length, whether the class met early or late in the day, and whether evaluations were administered online or on paper (see, for example, Isley and Singh [2005], McPherson [2006], Matos-Díaz and Ragan [2010], and Kherfi [2011]).[18] One of our key variables of interest is the prior level of student interest in the course on a scale of 1 to 5, with 5 being the greatest prior interest.

Another category of controls are faculty-semester specific variables that capture the number of students taught by the instructor. Class size is used in many papers analyzing the factors that affect SETs; we gathered this data from administrative records that list the actual number of students enrolled in the section after the drop/add period in the second week of a fifteen-week semester. We also include three variables that measure the instructor's overall teaching load: the total number of students that each instructor had enrolled in all of his or her class-sections that semester (student load), the number of course-sections taught by the instructor (course load), and the number of different courses that the instructor must prepare for in a given semester (number of preps).

The final set of explanatory variables includes semester, instructor, and course fixed effects. The semester fixed effects control for institutional change and observed evaluation inflation over time. The course fixed effects control for characteristics such as department, whether the course is required or elective, and whether the course is introductory or advanced.

Finally, instructor fixed effects control for any unobserved or unrecorded instructor characteristics. In particular, instructor fixed effects include characteristics related to the quality of teaching, such as how engaging the instructor is in class, how available outside of class, how clear in her explanations, how well-prepared, and so forth. In this we follow, for example, Rothstein (2010) and Goldhaber and Hansen (2010), both of which use teacher fixed effects to measure teacher value-added or effectiveness.[19] In other words, instructor fixed effects are a proxy for that part of the quality of instruction that does not vary over time. Naturally, we do not claim that instructor fixed effects are a precise measure of instructor quality. Because other characteristics that do not change over time, such as the instructor's sex, race, and accent, are subsumed in the fixed effects, if evaluation ratings are biased by any of these characteristics then the fixed effects will also be biased. In addition, the instructor fixed effects are an average over the time period of the study and therefore do not reflect any improvements in teaching effectiveness that may be the result of adopting new teaching techniques or attending teaching workshops or seminars. Because multiple instructors teach the same course, and (almost all) instructors teach multiple courses, the conditional faculty fixed effects eliminate the influence of course, section, and semester characteristics on evaluations and as such provide context and additional information to supplement the raw SET scores typically used to evaluate teaching. As described above, the econometric specification estimated is:

$$Y_{\mathrm{isjt}} = \beta_0 + \beta_1(X_{\mathrm{isjt}}) + \beta_2(Z_{\mathrm{it}}) + \delta_{\mathrm{i}} + \alpha_{\mathrm{j}} + \nu_{\mathrm{t}} + \varepsilon_{\mathrm{isjt}},$$

where Y represents the average course-section outcomes outlined above for instructor i in section s of course j at time (semester) t. The vector X includes the course-section-semester specific variables, and the vector Z reflects the instructor-semester specific variables discussed above. The error components δ, α, and ν represent instructor, course, and time effects, respectively, and standard errors are clustered by faculty member.[20] The primary focus of this analysis is on the estimated δ_i effects, which reflect the estimated faculty fixed effects conditional on the observable variables

outlined above, and the degree to which these conditional faculty fixed effects are correlated with the baseline faculty fixed effects (average faculty SET scores controlling just for semester effects).[21] This technique of using individual fixed effects as an estimate of worker ability conditional on workplace characteristics follows the approach in Abowd, Kramarz, and Margolis (1999), as applied, for example, in Cornelissen, Dustmann, and Schönberg (2017) and Card, Heining, and Kline (2013).

Table 1 presents summary measures of the data used in this analysis as well as those of additional questions on the SET instrument. It is interesting to note that the average class size at this school is only 23.39, and the average student load (total number of students taught by a single professor that semester) is 63.46, reflecting the relatively small size of this school. Less than 10 percent of classes were held early (defined as starting before 9:00 am), while on average 60 percent of the students in a class were male. In terms of the primary dependent variables of interest, the average response to the prompt, "I have learned a lot as a result of this course," is 4.12, with a range of 2.45 to 5.00 (the highest possible rating) and a standard deviation of 0.45; a rating of 5 indicates strong agreement, with 1 indicating strong disagreement, so that higher ratings are better. The statistics are similar for the SET questionnaire item, "this instructor's overall teaching ability is excellent": the average score is 4.18, with a range of 2.08 to 5.00 and a standard deviation of 0.48, using the same 1 to 5 Likert scale as above.

Estimation results

The baseline faculty fixed effects, conditional only on semester effects, are simply the coefficients on the set of instructor dummy variables relative to the instructor closest to the average rating on the "amount learned" SET questionnaire item. These coefficients vary from a high of almost 0.7 higher than the average faculty member to a low of 0.65 below the average faculty member. Faculty who are "noticeably" above the average are usually labeled as successful in teaching their students, while those below are usually identified as less effective at teaching their students.

The question addressed here is how well these baseline faculty scores correlate with faculty scores when controlling for course-section characteristics, faculty-semester characteristics, and course fixed effects. To clarify: we do not address how well SET scores correlate with actual teaching effectiveness, as we do not claim that our modified measure accurately gauges teaching effectiveness. Instead, our analysis addresses how well the measure that is used by many institutions (raw SET scores) is correlated with a measure that controls for multiple external factors outside the instructor's control. We first regress, using OLS and clustered (by faculty) standard errors, the average amount learned score for each course-section against all observable course-section and faculty-semester characteristics: class size, student load, course load, number of preps, meeting time (early or late), class duration (50 minutes versus 75 minutes), whether the evaluations were administered online, the proportion of the class that is male, the interaction of this variable with the instructor's sex, the cohort mix of the course-section, the average GPA of the students in the section, the average prior level of interest, whether the course was required or not, and semester and faculty fixed effects. Column 1 of table 2 presents the results; the estimated semester and faculty fixed effects are not individually reported in the table. The R-square value on this regression is 0.607, indicating that these regressors explain over half of the variation in average reported amount learned scores across course-sections.

It is not surprising that the larger the class, the lower was the average amount learned as reported by the students: on average, each additional student reduces the average rating of amount learned by 0.009.[22] The estimated coefficients for student load (number of students taught by the instructor in a given semester) and number of course preps are both negative, but neither is statistically significant. Similarly, although the number of course sections taught has a positive coefficient, it is also not statistically significant at conventional levels.[23]

Table 2. Regression results.

	Amount learned		Amount learned		Instructor rating		Instructor rating	
Intercept	3.835***		2.672***		4.404***		3.243***	
Class size	−0.009***	(0.0027)	−0.010***	(0.0025)	−0.008**	(0.003)	−0.009***	(0.003)
Student load	−0.001	(0.0001)	−0.002	(0.0014)	−0.001	(0.0014)	−0.001	(0.0016)
No. sections taught	0.039	(0.028)	0.043	(0.027)	0.027	(0.033)	0.042	(0.038)
No. of preps	−0.030	(0.026)	−0.026	(0.021)	−0.030	(0.028)	−0.034	(0.024)
Early	−0.095***	(0.026)	−0.072***	(0.025)	−0.117***	(0.030)	−0.083***	(0.029)
Late	0.009	(0.018)	0.007	(0.018)	0.013	(0.023)	0.004	(0.023)
50-minute class	0.051*	(0.029)	0.042	(0.026)	0.050*	(0.029)	0.036	(0.030)
Online evaluation	−0.010	(0.081)	0.024	(0.078)	−0.081	(0.087)	−0.024	(0.089)
Proportion male students	−0.093	(0.088)	−0.065	(0.054)	−0.216**	(0.084)	−0.235**	(0.072)
Male student * male faculty	0.022	(0.088)	0.030	(0.067)	0.198**	(0.096)	0.236***	(0.095)
Proportion: Sophomore	−0.003	(0.093)	0.168**	(0.070)	−0.025	(0.134)	0.228**	(0.101)
Junior	−0.008	(0.089)	0.365***	(0.093)	−0.018	(0.132)	0.492***	(0.141)
Senior	−0.120	(0.093)	0.324***	(0.110)	0.003	(0.130)	0.576***	(0.154)
GPA	−0.215**	(0.087)	−0.145*	(0.076)	−0.218**	(0.106)	−0.182**	(0.078)
Prior interest	0.317***	(0.041)	0.411***	(0.029)	0.186***	(0.042)	0.355***	(0.044)
Required course	−0.150***	(0.038)			−0.152***	(0.043)		
Semester fixed effects	Yes		Yes		Yes		Yes	
Faculty fixed effects	Yes		Yes		Yes		Yes	
Course fixed effects	No		Yes		No		Yes	
R^2	0.607		0.671		0.534		0.602	
Adjusted R^2	0.588		0.639		0.511		0.563	

Note: Standard errors are shown in parentheses.
*indicates two-tailed significance at the 90% level; **at 95%; and ***at 99%.

The coefficient estimate for early class-sections is negative, indicating that early classes (before 9:00 am) reduce the average reported amount learned by 0.095 relative to classes that meet later in the day. More frequent (3 versus 2 times a week) and thus shorter (50 versus 75 minutes) course-sections, on the other hand, have significantly higher average amount learned scores, although again the economic magnitude is small, increasing the SET rating by just 0.05. The impact of online evaluations, more male students, or more upperclassmen on the amount learned score is not statistically significant. Consistent with the results in Boex (2000) and Grimes, Millea, and Woodruff (2004), course-sections with higher average reported GPAs prior to the current course have lower reported average amount learned results: an increase of 0.1 in the average GPA reduces the average SET score by 0.02. Because the GPA measures students' performance prior to the semester in question and does not include the grade for the course being evaluated, this suggests that, on average, better students appear to be tougher graders of their instructors.[24]

The estimated coefficient on the average level of prior student interest in the course, one of our key variables of interest, is positive and significantly different from zero at the 99% level. The magnitude of the estimate indicates that a one unit increase in prior interest (on a 5-point scale) would increase the predicted average amount learned rating by over 0.3, approximately a 0.75 standard deviation increase in the amount learned.[25]

We also include a binary variable that controls for whether the instructor is teaching a required course or an elective, another key variable of interest. A required course is defined here as one required of all business students, regardless of their major or concentration. The estimation results indicate that the average rating of amount learned for a required course is 0.15 lower on average than for electives, and this estimate is again significant at the 99% level.

The correlation coefficient of the baseline and conditional faculty fixed effects from this regression is 0.81. This is the equivalent of an R-square value of 0.656 in a regression of the conditional fixed effects on the baseline faculty fixed effects.

In our second estimation, we include a full set of course dummy variables. Results are reported in column 2 of table 2.[26] As the courses are simply numbered 1 to 88, we are unable to categorize

or describe which types of courses have a lower or higher average amount learned. Nonetheless, incorporating course fixed effects into the regression raises the R-square value from 0.607 to 0.671 and lowers the correlation coefficient between the baseline and conditional faculty fixed effects to 0.59, with a Spearman's rank correlation coefficient of 0.50. A 0.59 correlation coefficient translates to an R-square value of only 0.34 in a regression of the conditional fixed effects on the baseline fixed effects. Clearly, for some faculty, a significant portion of the perceived amount learned depends on course and section attributes.[27]

The coefficient estimates for most of the control variables remain stable when course fixed effects are included (and the binary variable for required courses dropped). There are, however, a few notable differences in coefficient estimates. In particular, the coefficients on the proportion of students that are sophomores, juniors, and seniors are all positive and statistically significant. On average, sophomores within a given course report an amount learned score that is higher by 0.17 than that reported by first-years (the excluded category). Juniors and seniors report scores that are higher by 0.37 and 0.32, respectively.[28] The magnitude of the prior interest coefficient also increases substantially to 0.411, indicating that a 1-point increase in the prior interest value will increase amount learned scores by more than 0.4, a difference of almost one standard deviation. In addition, the coefficient on the class duration variable is still positive but no longer statistically significant. If faculty are allocated to teach certain courses, such as introductory or advanced seminar courses, based on teaching ability, then identification of separate faculty and course fixed effects becomes problematic and less precise.

We find that 23 faculty have estimated fixed effects that are statistically significantly different from zero, at the 95% level of confidence or higher. Additionally, the standard deviation of faculty fixed effect estimates is over twice the average of the standard errors of those estimates (standard deviation of faculty fixed effects = 0.2823; average standard errors = 0.1277). Clearly, there is significant variation in the estimated faculty fixed effects, relative to the standard errors of those estimates.

Including the course fixed effects produces estimates of faculty fixed effects that have a correlation coefficient of only 0.59 with the baseline faculty fixed effects that are produced controlling only for semester effects. Table 3a shows the quartiles of faculty ratings based on amount learned in both the baseline and conditional faculty fixed effects. Thus, for example, one faculty member who is in the fourth (lowest) quartile of the baseline fixed effects moves into the first (top) quartile based on the conditional faculty fixed effects, two faculty in the bottom baseline fixed effects quartile move up into the second conditional fixed effects quartile, and so forth.

If the two approaches produced identical quartile groupings of faculty, then we would expect to see only the diagonal cells of the table populated. In fact, many faculty move between quartile rankings once course and semester controls are incorporated. Specifically, 19 out of 45 faculty (42%) remain in the same quartile, while 16 (36%) move one quartile, 8 (18%) move two quartiles, and two instructors (4%) move three quartiles. This indicates that course attributes have a significant influence on SET ratings, and because multiple faculty teach each course these attributes are related to the content of the course rather than to the instruction itself. For example, the

Table 3a. Amount learned quartile rankings.

		Baseline faculty fixed effects				
		1st quartile	2nd quartile	3rd quartile	4th quartile	Total
Conditional faculty fixed effects	1st quartile	6	2	2	1	11
	2nd quartile	3	4	2	2	11
	3rd quartile	1	2	4	4	11
	4th quartile	1	3	3	5	12
	Total	11	11	11	12	45

impact of course attributes could depend on whether a course is heavily quantitative, whether a course is abstract or concrete, or whether course readings are highly technical. Because individual courses are not identified in the dataset, we cannot determine which attributes in particular tend to be associated with higher SET ratings. We can, however, conclude that SET ratings depend in no small part on the course being taught. These course-specific effects suggest that SET scores and faculty ratings need to be assessed within the context of the instructors' assigned courses.

The effectiveness of an instructor is often assessed relative to the mean or median ratings in a given institution. In our analysis, controlling for course, section, and semester attributes (i.e., using the fully conditional faculty fixed effects) moves 7 faculty (16%) from below the median rating to above the median, 2 of whom are significantly below the median at the 95% level of confidence. Naturally, 7 faculty move from above to below the median rating, 2 of whom are significantly above the median at the 95% level of confidence. Relative to the mean, 17 faculty move from below to above the mean (6 of whom are below the mean with a 95% level of confidence), while 4 faculty move in the opposite direction (1 of whom is above the mean with a 95% level of confidence). These lopsided shifts result from the fact that the baseline fixed effects are right-skewed while the fully conditional fixed effects are left-skewed.

While these statistics clearly indicate that a substantial number of faculty are significantly impacted, it is also worth noting that the fully conditional fixed effects drastically change the ratings of a few individual faculty members, moving them from being among the lowest rated to being about average.[29] Overall, the average of the absolute value of the differences between the baseline and conditional faculty fixed effects is 0.234, which is more than one-half of a standard deviation in the average rating across all sections and instructors.[30]

We repeated the above analysis using the overall instructor quality ratings in place of the amount learned scores. Columns 3 and 4 of table 2 repeat, using this overall instructor quality score, the analyses in columns 1 and 2: column 3 controls for course-section and faculty-semester attributes, including prior interest and whether a course is required, and column 4 includes a full set of course fixed effects. The coefficient estimates in column 3 are qualitatively the same in terms of sign and significance as those in column 1, with two exceptions. First, the estimate of the coefficient on the proportion of male students is statistically significant at the 95% level: male students give lower SET scores overall. The estimate on the interaction of proportion of male students with the instructor's sex is also significant at the 95% level but indicates that males give higher SET scores to male faculty. Given the average class size of just over 23, one additional male student would lower the average SET rating of the overall quality of a female instructor by just less than 0.01 and raise it for a male instructor by approximately the same amount.

The remaining coefficient estimates are quite similar to those in column 1. Once again, larger class size, earlier classes, classes of longer duration, and students with higher GPAs lower the instructor quality scores. The students' level of prior interest coefficient estimate is positive and significantly different from zero at the 99% level, as in column 1. The magnitude of the coefficient is notably smaller than in column 1 and indicates that a one-unit increase in student prior interest in the subject is correlated with a 0.186 increase in the overall instructor rating. The estimate of the effect of a required course on ratings of overall teaching quality, also statistically significant at the 99% level, is remarkably similar to that in column 1: the coefficient estimate indicates that required courses have overall instructor ratings that are 0.152 lower, on average, than electives, conditional on the other variables. The R-square value on this regression is 0.534, indicating that approximately one half of the variation in the overall instructor quality rating is explained by the course-section and faculty-semester attributes as well as semester and faculty fixed effects.

Column 4 of table 2 includes the full set of course fixed effects for the overall instructor quality score. Coefficient estimates are the same in sign and significance as those in column 2, with the exception of the proportion of male students and the interaction of this variable with the

Table 3b. Instructor rating quartile rankings.

		Baseline faculty fixed effects				
		1st quartile	2nd quartile	3rd quartile	4th quartile	Total
Conditional fixed effects	1st quartile	6	3	2	0	11
	2nd quartile	4	5	1	1	11
	3rd quartile	1	3	4	3	11
	4th quartile	0	0	4	8	12
	Total	11	11	11	12	45

male faculty indicator (both of which are now statistically significant) and the online evaluation indicator (which changes sign but is still not statistically significant). The magnitude of the coefficients is also significantly larger for the proportion of male students, the proportion of male students interacted with male faculty indicator, and the share of each section that is sophomores, juniors, and seniors.

The correlation coefficient between the baseline faculty fixed effects (controlling just for semester effects) and the fully conditional faculty fixed effects is 0.79. This is equivalent to an R-square value of 0.62 in a regression of the fully conditional faculty fixed effects on the baseline fixed effects. In other words, 62 percent of the variation in faculty fixed effects can be predicted by the baseline effects, while 38 percent of the variation is unexplained by the baseline evaluation. Twenty-six of the faculty fixed effects are statistically significantly different from zero at the 95% level, or higher. The standard deviation of faculty fixed effects estimates in this case is over three times greater than the average standard errors of those estimates (standard deviation of the faculty fixed effects = 0.365; average standard errors of the estimated faculty fixed effects = 0.116). For this survey item, there is even greater variation in the estimated faculty fixed effects, relative to the standard errors of those estimates.

Table 3b shows that 23 out of 45 (51%) faculty would remain in the same quartile ranking based on fully conditional and baseline faculty fixed effects. Eighteen faculty members (40%) would move one quartile, and another four (9%) would move two quartiles based on conditional faculty fixed effects. Nearly one-fifth of the faculty in our sample (18%) move either from below to above the median or the reverse, although only two out of these eight faculty are significantly different from the median value, at the 95% level of confidence. And, eight faculty move from below the mean to above the mean, with three of these faculty significantly below the mean at the 95% level of confidence or higher, while two move from above to below the mean, although neither of the individuals is significantly above the mean.

The average absolute value of the difference between the baseline and conditional faculty fixed effects is 0.199 (close to one-half of a standard deviation). Thus, as with the student ratings of amount learned, the ratings of overall instructor quality are markedly different when controlling for course-section and faculty-semester attributes and particularly course fixed effects.

Conclusions

Many papers have analyzed the effects of various course, section, and faculty attributes on student evaluations of teaching results. We use a dataset covering 48 faculty members, 88 courses, 26 semesters, and 1,928 observations to study the impact of a wide range of course-section and faculty-semester attributes on SET ratings. We focus especially on how students' desire to take a class may impact their SET ratings and find that SET scores are higher when students' level of interest prior to taking the course is higher and when the course is an elective rather than a required course. We then take the analysis a step further and consider not only how these factors affect SET ratings but also how this can in turn affect the evaluation of faculty members' teaching effectiveness. Comparing the baseline faculty fixed effects with only semester controls to the fully

conditional faculty fixed effects, including required and prior interest variables, reveals significant differences in ratings for a majority of the faculty. When we include course fixed effects as well as the students' prior level of interest variable, we find even greater differences in both the SET scores and the relative faculty rankings.

These results indicate that the identification of effective teaching based on SET ratings of overall instructor quality and amount learned, as is a common practice in many institutions, will be markedly different when controlling for course fixed effects, students' prior level of interest, and a series of other variables that characterize each section and each instructor's teaching load. Of course, the fully conditional faculty fixed effects absorb a number of time-invariant factors, some of which are related to teaching effectiveness (e.g., how engaging is the instructor, does the instructor respect students, is the instructor on time and prepared) and some of which are not (e.g., race, sex, accent). Furthermore, they do not account for changes in teaching effectiveness over time. As such, the conditional fixed effect estimates should not be taken as a precise measure of teaching effectiveness but instead as a rough proxy of those elements of teaching effectiveness that are constant over time. Indeed, the substantial impact that prior student interest variables and other control variables have on SET scores suggests that SET scores should be interpreted in the context of the courses being taught, as they may be influenced by many factors unrelated to teaching effectiveness. This is especially so when the rankings cross departments and divisions, given how much course fixed effects matter to the rankings. A ranking based on fully conditional faculty fixed effects can provide such context for assessing instructional performance. Future research attempting to assess teaching quality may find tracking student performance across multiple instructors, in the manner of Carrell and West (2010), a fruitful avenue of investigation, with Rothstein's (2010) analysis providing a strict structural framework for undertaking this type of study. Analyzing data across schools within an institution or even across institutions can provide additional insight into the factors that increase teaching effectiveness. Such an examination would allow controlling for both observed and unobserved differences across instructors (see Abowd, Kramarz, and Margolis 1999). It would also allow identification of peer effects (as in Cornelissen, Dustmann, and Schönberg 2017) or school or institutional characteristics that impact teaching (similar to Card, Heining, and Kline 2013), shedding light more broadly on what makes for effective teaching.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes

1. Many institutions use a variety of approaches to measure teaching effectiveness, from SETs to classroom observation to review of teaching and assessment materials. However, Becker and Watts (1999) report that the use of SETs is almost universal (of 302 U.S. institutions responding to their survey, only three did not use SETs). Becker, Bosshardt, and Watts (2012) confirm the prevalence of SETs and note further that the average weight across all types of institutions that is given to SETs in evaluating teaching effectiveness is 48.7%.
2. Examples of the summary question are "Overall, how would you rate the teaching effectiveness of this instructor," "Overall, how would you rate this course," and "Would you recommend this class to another student?" Most SETs include at least one such question.
3. Semester effects account for any "evaluation inflation" that is present. Although most institutions do not explicitly estimate semester fixed effects, this is approximated by comparing SET scores semester by semester, as many institutions do.
4. It is important to note that we do not claim that the conditional faculty fixed effects that control for multiple factors are a comprehensive measure of teaching effectiveness, but rather that they provide important additional information that may be used in assessing teaching performance.

5. See Ongeri (2009) for a discussion of the characteristics and content of economics courses that may result in negative student evaluations.
6. Al-Issa and Sulieman (2007) identify 2,988 papers that focus on student evaluations of teaching and were published in scholarly journals between 1990 and 2005.
7. It is not surprising that instructor experience has a significant and positive impact on SET ratings. Indeed, if experience helps build human capital, then it may be appropriate to consider experience as a component of teaching effectiveness.
8. The mean principles class size in McPherson's (2006) data is 82 for principles classes, while the mean class size for upper-level classes is 33. Class sizes in our dataset are no larger than 45 and most are substantially smaller.
9. For example, if SET ratings are substantially lower for required courses, then an instructor's ranking *within* a subset of courses that are all required (e.g., principles courses) may not change when course fixed effects are included. However, when all courses are pooled together, the principles instructors' rankings may improve relative to the nonprinciples instructors' when course fixed effects (which include whether a course is required) are factored in.
10. Given our result that required courses (e.g., principles) receive lower ratings on average, it is possible that faculty who teach principles courses try to compensate for this trend and therefore, conditional on course attributes, are actually more effective instructors.
11. We note that Carrell and West (2010) do not argue that their approach is one that should be broadly implemented.
12. While a number of other control variables are subsumed by the faculty and course-semester random effects that Carrell and West (2010) include in their estimation, these three attributes vary within either the course-semester or the faculty identifier; they are therefore not accounted for by the random effects and instead enter into the error term.
13. Content validity is suspect both because no precise definition of all the elements of teaching effectiveness exists at most institutions, and students (who complete the SETs) and faculty or administrators (who design them) seem to disagree about what constitutes effective teaching. Construct validity is called into question because, for example, SETs do not effectively distinguish between teaching effectiveness and how much students like the instructor; see Onwuegbuzie, Daniel, and Collins (2009), Benton and Cashin (2012), and Clayson (2015) for further discussion.
14. Naturally, we do not include any variables that are specific to the instructor but do not vary over time, as these are subsumed by the faculty fixed effects.
15. Their approach is applicable only in those institutions where ratings are calculated by averaging an instructor's scores across not only multiple students but also multiple questions on the instrument.
16. A course-section is defined by a combination of course, meeting time, and instructor.
17. See Sproule (2002) for theory underlying the inclusion of these variables.
18. The GPA gives the average of students' self-reported GPA across all courses prior to that semester. The gender mix of the section comprises two variables: the proportion of the students who were male, and an interaction between this proportion and a binary variable equal to one if the instructor was male. The grade level mix of the section comprises three variables: the proportion of students who were sophomores, juniors, and seniors (first-years was the excluded category). Class length is a binary variable equal to one if the class met three times a week for 50 minutes; otherwise, the class met twice a week for 75 minutes each. Class time comprises two variables: the first is a binary variable equal to one if the class met before 9:00 am, the second is a binary variable equal to one if the class met after 2:00 pm, and the excluded meeting times are those in the middle of the day.
19. Note that Rothstein (2010) refers to "teacher" and "classroom" effects interchangeably because teachers are assigned to a single classroom in his data on elementary students in North Carolina. For a survey of value-added estimation, see McCaffrey et al. (2004).
20. We also estimate equation (1) using weighted least squares, with weights equal to the number of student responses in each section. Results are qualitatively similar and are available from us upon request.
21. Of course, omitting factors that are significant in explaining variation in the SET ratings only results in biased coefficient estimates if the excluded variables are correlated with the included variables. As Greene (2000, 229) explicitly states, "If the variables in a multiple regression are not correlated (i.e. are orthogonal), then the multiple regression slopes are the same as the slopes in the individual simple regression." This property of multiple regression estimates implies that a simple comparison of raw SET averages across faculty members may in fact be an accurate estimate of teaching effectiveness, as long as the course and faculty characteristics that are significant in explaining variation in the raw SET scores (which are by definition not conditional) are not correlated with the (unobserved) teaching effectiveness of the instructors teaching those courses.

22. McPherson (2006) and Ragan and Walia (2010) both find that increased class size leads to lower ratings; Ragan and Walia also find that the effect decreases with class size.
23. The negative relationship between class size and perceived learning may be due to focusing less attention on each student during class time. Negative student load and course prep coefficients could easily be explained by faculty having less time to spend providing individual assistance to students outside of class or improving the quality of any individual course. A positive coefficient for sections taught may be attributable to faculty devoting more attention and effort to their teaching during semesters when they have a higher teaching load and focusing more on research and other activities when they have fewer teaching responsibilities.
24. It is, of course, also possible that instructors try to incentivize students to give high SET ratings with easy courses in which students received high grades but do not learn a lot. However, because the GPA is the average of students' GPAs across all prior classes and does not include the class being evaluated, the grade inflation would have to be present in all courses across the entire university in order for a high GPA to indicate easy courses rather than better students.
25. While cognitive dissonance may be a problem in this instance due to asking students at the end of the semester to recall their level of interest in the subject prior to taking the class, any bias resulting from this would actually be likely to result in under-estimating the relationship between prior interest and amount learned. Students who do not enjoy a course may well perceive both that they have learned little and that the instructor has diminished their interest in the subject. In this case, the prior interest rating will be biased upward but the amount learned rating will be biased downward, implying that the coefficient is actually under-estimated in our results.
26. Naturally, this requires dropping the binary variable that indicates a required course. Additionally, controlling for course fixed effects eliminates two faculty from the estimated faculty fixed effects due to perfect collinearity, resulting in 45 faculty fixed effects estimated, relative to the omitted average faculty member.
27. A significant portion of this is due to including the required course and prior interest variables. When faculty fixed effects are estimated without controlling for these two variables, the correlation with the baseline faculty fixed effects is 0.98.
28. This could be due to a handful of introductory courses, populated primarily with first-year students, receiving higher than average scores in amount learned.
29. The overall results are not driven by these individual faculty members.
30. The differences are somewhat skewed: just under 25 percent of faculty included in the sample receive unconditional ratings that are higher than the fully conditional ratings, indicating that a portion of their high unconditional ratings are actually due to course fixed effects or course-section or faculty-semester characteristics, e.g., small class size or student load, students with high GPAs, etc. But three quarters (76%) of faculty receive unconditional ratings that are lower than the fully conditional ratings.

References

Abowd, J., F. Kramarz, and D. Margolis. 1999. High wage workers and high wage firms. *Econometrica* 67 (2): 251–333.

Al-Issa, A., and H. Suleiman. 2007. Student evaluations of teaching: Perceptions and biasing factors. *Quality Assurance in Education* 15 (3): 302–17.

Baldwin, T., and N. Blattner. 2003. Guarding against potential bias in student evaluations: What every faculty member needs to know. *College Teaching* 51 (1): 27–32.

Becker, W. E., W. Bosshardt, and M. Watts. 2012. How departments of economics evaluate teaching. *Journal of Economic Education* 43 (3): 325–33.

Becker, W. E., and M. Watts. 1999. How departments of economics evaluate teaching. *American Economic Review: Papers and Proceedings* 89 (2): 344–49.

Benton, S. L., and W. E. Cashin. 2012. Student ratings of teaching: A summary of research and literature. IDEA Paper #50. Manhattan, KS: IDEA. https://www.ideaedu.org/Portals/0/Uploads/Documents/IDEA%20Papers/IDEA%20Papers/PaperIDEA_50.pdf.

Boex, L. F. J. 2000. Attributes of effective economics instructors: An analysis of student evaluations. *Journal of Economic Education* 31 (3): 211–27.

Boring, A. 2017. Gender biases in student evaluations of teaching. *Journal of Public Economics* 145:27–41.

Boring, A., K. Ottoboni, and P. B. Stark. 2016. Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*. doi: 10.14293/s2199-1006.1.sor-edu.aetbzc.v1

Braga, M., M. Paccagnella, and M. Pellizzari. 2014. Evaluating students' evaluations of professors. *Economics of Education Review* 41:71–88.

Campbell, H. E., S. Steiner, and K. Gerdes. 2005. Student evaluations of teaching: How you teach and who you are. *Journal of Public Affairs Education* 11 (3): 211–31.
Card, D., J. Heining, and P. Kline. 2013. Workplace heterogeneity and the rise of West German wage inequality. *Quarterly Journal of Economics* 128 (3): 967–1015.
Carrell, S. E., and J. E. West. 2010. Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy* 118 (3): 409–32.
Clayson, D. E. 2015. A multi-disciplined review of the student teacher evaluation process. Mimeo. Cedar Falls, IA: University of Northern Iowa, College of Business Administration. http://business.uni.edu/clayson/set.
Cornelissen, T., C. Dustmann, and U. Schönberg. 2017. Peer effects in the workplace. *American Economic Review* 107 (2): 425–45.
d'Apollonia, S., and P. C. Abrami. 1997. Navigating student ratings of instructions. *American Psychologist* 52 (11): 1198–1208.
De Witte, K., and N. Rogge. 2011. Accounting for exogenous influences in performance evaluations of teachers. *Economics of Education Review* 30:641–53.
Dilts, D. A. 1980. A statistical interpretation of student evaluation feedback. *Journal of Economic Education* 11 (2): 10–15.
Driscoll, J., and D. Cadden. 2010. Student evaluation instruments: The interactive impact of course requirement, student level, department and anticipated grade. *American Journal of Business Education (AJBE)* 3 (5): 21–30.
Ewing, A. M. 2012. Estimating the impact of relative expected grade on student evaluations of teachers. *Economics of Education Review* 31 (1): 141–54.
Feldman, K. A. 1993. College students' views of male and female college teachers: Part II: Evidence from students' evaluations of their classroom teachers. *Research in Higher Education* 34 (2): 151–211.
_____. 2007. Identifying exemplary teachers and teaching: Evidence from student ratings. In *The scholarship of teaching and learning in higher education: An evidence-based perspective*, ed. R. P. Perry and J. C. Smart, 93–143. The Netherlands: Springer Science & Business Media.
Fenn, A. J. 2015. Student evaluation based indicators of teaching excellence from a highly selective liberal arts college. *International Review of Economics Education* 18:11–24.
Goldhaber, D., and M. Hansen. 2010. Using performance on the job to inform teacher tenure decisions. *American Economic Review* 100 (2): 250–55.
Gramlich, E. M., and G. A. Greenlee. 1993. Measuring teaching performance. *Journal of Economic Education* 24 (1): 3–13.
Greene, W. H. 2000. *Econometric analysis*. 4th ed. Upper Saddle River, NJ: Prentice Hall.
Grimes, P. W., M. J. Millea, and T. W. Woodruff. 2004. Grades—Who's to blame? Student evaluation of teaching and locus of control. *Journal of Economic Education* 35 (2): 129–47.
Hamermesh, D. S., and A. Parker. 2005. Beauty in the classroom: Instructors' pulchritude and putative pedagogical productivity. *Economics of Education Review* 24 (4): 369–76.
Isley, P., and H. Singh. 2005. Do higher grades lead to favorable student evaluations? *Journal of Economic Education* 36 (1): 29–42.
Kherfi, S. 2011. Whose opinion is it anyway? Determinants of participation in student evaluation of teaching. *Journal of Economic Education* 42 (1): 19–30.
Krautmann, A. C., and W. Sander. 1999. Grades and student evaluations of teachers. *Economics of Education Review* 18 (1): 59–63.
Langbein, L. 2008. Management by results: Student evaluation of faculty teaching and the mis-measurement of performance. *Economics of Education Review* 27 (4): 417–42.
Marks, R. B. 2000. Determinants of student evaluations of global measures of instructor and course value. *Journal of Marketing Education* 22 (2): 108–19.
Matos-Díaz, H., and J. F. Ragan, Jr. 2010. Do student evaluations of teaching depend on the distribution of expected grade? *Education Economics* 18 (3): 317–30.
McCaffrey, D. F., J. R. Lockwood, D. Koretz, T. A. Louis, and L. Hamilton. 2004. Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics* 29 (1): 67–101.
McPherson, M. A. 2006. Determinants of how students evaluate teachers. *Journal of Economic Education* 37 (1): 3–20.
McPherson, M. A., and R. T. Jewell. 2007. Leveling the playing field: Should student evaluation scores be adjusted? *Social Science Quarterly* 88 (3): 868–81.
McPherson, M. A., R. T. Jewell, and M. Kim. 2009. What determines student evaluation scores? A random effects analysis of undergraduate economics classes. *Eastern Economic Journal* 35 (1): 37–51.
Monks, J., and R. M. Schmidt. 2011. The impact of class size on outcomes in higher education. *BE Journal of Economic Analysis & Policy* 11 (1).
Olivares, O. J. 2001. Student interest, grading leniency, and teacher ratings: A conceptual analysis. *Contemporary Educational Psychology* 26 (3): 382–99.

Ongeri, J. D. 2009. Poor student evaluation of teaching in economics: A critical survey of the literature. *Australasian Journal of Economics Education* 6 (2): 1–24.

Onwuegbuzie, A. J., L. G. Daniel, and K. M. Collins. 2009. A meta-validation model for assessing the score-validity of student teaching evaluations. *Quality & Quantity* 43 (2): 197–209.

Onwuegbuzie, A. J., A. E. Witcher, K. M. T. Collins, J. D. Filer, C. D. Wiedmaier, and C. W. Moore. 2007. Students' perceptions of characteristics of effective college teachers: A validity study of a teaching evaluation form using a mixed-methods analysis. *American Educational Research Journal* 44 (1): 113–60.

Ragan, J. F., and B. Walia. 2010. Differences in student evaluations of principles and other economics courses and the allocation of faculty across courses. *Journal of Economic Education* 41 (4): 335–52.

Rothstein, J. 2010. Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics* 125 (1): 175–214.

Spooren, P., B. Brockx, and D. Mortelmans. 2013. On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research* 83 (4): 598–642.

Sproule, R. 2002. The underdetermination of instructor performance by data from the student evaluation of teaching. *Economics of Education Review* 21 (3): 287–94.

Stark, P. B., and R. Freishtat. 2014. An evaluation of course evaluations. *ScienceOpen Research* (September 29). https://www.scienceopen.com/document?id=6233d2b3-269f-455a-ba6b-dc3bccf4b0a8.

Wagner, N., M. Rieger, and K. Voorvelt. 2016. Gender, ethnicity and teaching evaluations: Evidence from mixed teaching teams. *Economics of Education Review* 54:79–94.

Weinberg, B. A., M. Hashimoto, and B. M. Fleisher. 2009. Evaluating teaching in higher education. *Journal of Economic Education* 40 (3): 227–61.