

Measuring teaching quality in higher education: assessing selection bias in course evaluations

Maarten Goos¹ · Anna Salomons¹

Received: 12 February 2015 / Published online: 22 July 2016

© The Author(s) 2016. This article is published with open access at Springerlink.com

Abstract Student evaluations of teaching (SETs) are widely used to measure teaching quality in higher education and compare it across different courses, teachers, departments and institutions. Indeed, SETs are of increasing importance for teacher promotion decisions, student course selection, as well as for auditing practices demonstrating institutional performance. However, survey response is typically low, rendering these uses unwarranted if students who respond to the evaluation are not randomly selected along observed and unobserved dimensions. This paper is the first to fully quantify this problem by analyzing the direction and size of selection bias resulting from both observed and unobserved characteristics for over 3000 courses taught in a large European university. We find that course evaluations are upward biased, and that correcting for selection bias has non-negligible effects on the average evaluation score and on the evaluation-based ranking of courses. Moreover, this bias mostly derives from selection on unobserved characteristics, implying that correcting evaluation scores for observed factors such as student grades does not solve the problem. However, we find that adjusting for selection only has small impacts on the measured effects of observables on SETs, validating a large related literature which considers the observable determinants of evaluation scores without correcting for selection bias.

Keywords Education quality · Student evaluations of teaching (SET) · Sample selection bias · Heckman selection model

✉ Anna Salomons
a.m.salomons@uu.nl

Maarten Goos
m.goos@uu.nl

¹ Utrecht University School of Economics, P.O. Box 80125, 3508 TC Utrecht, The Netherlands

Introduction

Education quality matters for learning outcomes (Hanushek et al. 2005; Hanushek and Rivkin 2006) as well as outcomes later in life (Chetty et al. 2014b), and an important aspect of education quality is the quality of instruction (Hanushek and Rivkin 2006; Nye et al. 2004; Rivkin et al. 2005; Rockoff 2004). However, there is less agreement on the appropriate measurement of teaching quality, particularly for higher education where the availability of standardized testing to directly compare teaching performance is rare.¹ Yet public interest in this measurement is rising with the marked increase in college attainment in many countries, since institutions of higher education to a large extent rely on scarce public resources.

The most widespread measure of teaching quality in higher education are Student Evaluations of Teaching (SETs), also known as course evaluations.² Such course evaluations are used on a large scale to assess the quality of instruction at institutions of higher education as well as for comparing teacher performance across courses, departments and even universities (Becker and Watts 1999; Becker et al. 2011). Indeed, SETs affect faculty promotions (Galbraith et al. 2012; Kogan et al. 2010), students' college applications (Alter and Reback 2014) as well as students' course selection (Coleman and McKeachie 1981; Leventhal et al. 1975, 1976; Wilhelm 2004) and are used for published institutional teaching rankings³ and in auditing practices for demonstrating institutional performance (Johnson 2000). These practices have consequently sparked a large scientific literature on SETs, spanning across different fields such as education science, psychology, sociology, and economics.

The overarching question posed by these different scholars is to what extent SETs can be interpreted as an indicator for teaching quality in higher education. On the one hand, Marsh (2007) and Spooren et al. (2013) survey a large education literature and conclude that there is empirical, conceptual, and theoretical support for SETs. For example, this literature considers the psychometric properties of the evaluation questionnaires, assessing the extent to which scores generated by SETs measure the variables they intended to capture (Centra 1993).⁴ Furthermore, this work shows that SETs are positively correlated with teachers' self-evaluations and with ratings by trained observers.

¹ At the primary or secondary education level, the consensus is that teacher quality is best measured by means of “teacher value added”, essentially coefficients on teacher fixed effects in a regression with test scores as the dependent variable (e.g. see Chetty et al. 2014a, b), although there is some evidence that parents do not respond to these (Imberman and Lovenheim 2013; Pope 2014). These measures are not broadly applicable at the level of higher education, because tests are not standardized—an essential component for the measurement of teacher value added. See Cunha and Miller (2014) for an in-depth exploration of the possibilities and limitations of value-added measures in higher education.

² Other approaches such as peer evaluation and teacher certification also occur, but these are much less common and generally given little weight in assessments of teaching quality, unlike SETs which are both widely used and highly weighted (Becker and Watts 1999; Becker et al. 2011; Zabaleta 2007). Furthermore, there is little evidence that certified teachers are more effective (Angrist and Guryan 2008; Kane et al. 2008; Rivkin et al. 2005) even though teachers who are more qualified in the field they are teaching provide higher-quality instruction than so-called adjunct teachers (e.g. see Carrell and West 2010).

³ Whether officially published ones (e.g. the Performance Indicators Steering Group in the UK; the magazine Elsevier's Study Choice Guide in the Netherlands), or unofficial evaluations such as the US website ratemyprofessor.com.

⁴ However, this approach has also been criticized: Morley (2012) finds that measures of interrater agreement can be low despite high (intra-class) reliability coefficients.

On the other hand, the use of course evaluations as a measure of teaching quality has also been criticized for a number of reasons (Spooren et al. 2013). In particular, it is often argued that the signal SETs provide on teacher quality is contaminated by noise. Indeed, evaluation results tend to partially reflect student, course, and teacher characteristics which may not be related to teaching quality (e.g. see Berk 2005; Isely and Singh 2005; MacNeill et al. 2015; McPherson et al. 2009).⁵ Related to this is the argument that such noise in course evaluations provides teacher incentives for grade inflation, since there is evidence to suggest that the average (expected) grade has a positive effect on course evaluations irrespective of learning outcomes (Ewing 2012; Krautmann and Sander 1999; Langbein 2008; McPherson 2006). However, such concerns can be addressed in practice: as long as there is some true information contained in course evaluations about teaching quality, we can adjust for such observable course and teacher characteristics to filter out the noise, as suggested by Greenwald and Gillmore (1997), McPherson (2006), McPherson and Jewell (2007), and McPherson et al. (2009), among others. In particular, it has been suggested that grade differences should be removed to make SETs comparable across courses and remove incentives for grade inflation.

In this paper, we address a more fundamental concern arising from SETs as a measure of teaching quality: possible selection bias resulting from non-response. In most institutions, SET response rates are quite low, raising the concern that course evaluation scores are not representative of all students' opinions (Nulty 2008). After all, there will be selection bias if the students who choose to participate in the evaluation are not randomly selected, and its size and direction are important considerations for institutions wanting to measure student satisfaction with teaching as well as compare teaching quality across courses. Furthermore, research findings in the literature are potentially biased: selection may not only bias the average evaluation score, but also the coefficients of a regression aimed at uncovering the determinants of the evaluation score. Lastly, any such problems are exacerbated in online evaluations, which are increasingly widespread due to their cost-effectiveness relative to paper-based SETs (Becker and Watts 1999; Becker et al. 2011) but typically have significantly lower response rates (Avery et al. 2006; Dommeyer et al. 2004; Ho and Shapiro 2008; Liegle and McDonald 2004; Liu 2006; Nulty 2008; Shih and Fan 2009). For these reasons, we consider analyzing selection bias in SETs a first-order concern.

This question has remained largely unaddressed in the literature, possibly because data on non-respondents is typically unavailable. Nulty (2008) discusses the potential implications of low response rates in terms of selection bias, but also claims that it is impossible to empirically determine the direction of this bias. Furthermore, Adams and Umbach (2012), Kherfi (2011), and Spooren and Van Loon (2012) study the determinants of course evaluation response, comparing the observable characteristics of respondents to those of non-respondents. All three studies find significant differences: for example, respondents on

⁵ A more extreme version of the argument that course evaluations have a low signal-to-noise ratio (i.e. are contaminated by information unrelated to teaching quality) is the argument that course evaluations are *only* noise (i.e. are not related to teaching quality at all). However, there is as yet no consensus on this (Clayson 2009). For instance, Braga et al. (2014), Carrell and West (2010), and Weinberg et al. (2009) find that student evaluations are not correlated to learning, as measured by performance in follow-up courses. On the other hand, Hoffmann and Oreopoulos (2009) find that evaluations predict teacher quality as measured by drop-out rates, and Beleche et al. (2012) find a positive, albeit small, correlation between evaluations and subsequent student learning. Finally, Jacob and Lefgren (2008) conclude that SETs only identify the best and worst teachers based on achievement gains, while being much less accurate across the middle of the teaching quality distribution.

average have higher grades. Further, in an analysis of the determinants of SET scores, McPherson (2006) controls for the response rate at the course level to correct for selection in a reduced—form way, finding it to have a positive, albeit small, effect on upper-division courses. These studies provide suggestive evidence that SETs may be biased, but none of them examines to what extent such non-random response actually produces selection bias in SETs: doing so would require either performing a randomized experiment exogenously manipulating the response rate, or estimating a Heckman selection model (Heckman 1974).

There is some evidence from the experimental approach: in particular, Dommeyer et al. (2002, 2004) compare the results of paper and online SETs in a setting where students across 16 courses were randomly divided to one of these two modes of evaluation. Despite response rates being lower in the online evaluation, the authors do not find significantly different evaluation scores. It should be noted, however, that the response rate of students to the experimental questionnaire was itself only 62 %, implying information on non-responders was incomplete. While these results may be considered as *prima facie* evidence that lower response rates do not change the representativeness of SETs, this is far from conclusive: there is strong reason to suspect that the selection process differs between paper-based and online evaluations (Nulty 2008). Indeed, Dommeyer et al. (2002) show that the reasons students themselves provide for not responding differ between the paper-based and online modes.⁶ Given different selection processes, these results cannot be interpreted as reflecting the impact of different response rates on evaluation scores: nor do Dommeyer et al. (2002, 2004) make this claim—their studies are mostly focused on evaluating various incentive schemes to increase response.

Lastly, complementary work by Wolbring (2012) and Wolbring and Treischl (2015) considers selection bias in paper-based evaluations by examining the effects of class absenteeism on the day of the evaluation across courses. In particular, these studies find that selection bias due to observed characteristics is positive but quite small, although adjusting for it still has important effects on the ranking of courses. However, this work does not consider that selection bias may additionally arise from unobserved characteristics, nor does it inform about selection bias in the increasingly common online SETs, which have lower response rates and where class absenteeism does not constrain response to the evaluation questionnaire.

Our study goes beyond the existing approaches and findings in the literature in three main ways. Firstly, we explicitly quantify the selection bias (in terms of both sign and size) in SETs using a formal Heckman selection model which is empirically identified through an instrumental variable strategy. This implies we analyze selection on both observed and unobserved characteristics: this is particularly important since previous work has shown the limited explanatory power of observable factors over evaluation scores (Marsh 2007; Spooen et al. 2013). Furthermore, if selection bias is primarily a matter of selection on student grades or other factors that most university administrators can observe, it can easily be corrected. Correcting for selection on unobservables is a decidedly more difficult exercise, which is unlikely to be adopted by institutions on a large scale. The second contribution is that we are the first to consider selection bias in a web-based SET setting, which is the most common mode of evaluation in institutes of higher education. Our third contribution is considering the validity of a large literature which examines the

⁶ For example, non-response in online SETs came from forgetting the deadline, whereas for paper-based SET absence from class was the most important reason for not responding. Similarly, online SETs were perceived by students as less anonymous but more convenient as compared to paper-based SETs, suggesting that the cost-benefit structure of these two modes differs.

determinants of course evaluation scores: in the face of selection bias, the findings from this literature may be biased, as well. These contributions are made by using detailed information from a large European university. Our data spans more than 28,000 students taught in over 3000 courses across 14 different faculties, and includes information on all SET respondents and non-respondents.

The remainder of this paper is structured as follows. In the next section, we describe our data by discussing the institutional background and providing summary statistics. The third section then outlines the selection problem for course evaluations, covering both selection on observables and on unobservables. Results from our estimated selection model are documented and discussed in the fourth section. In the fifth section, we address the external validity of our results and outline a number of limitations. The final section concludes and offers policy implications based on our findings.

Data Description

Institutional Background

Our data hails from a large European university in the academic year 2010/2011. Course evaluation at this university is web-based: after each semester has been completed (i.e. teaching has concluded and exam grades have been announced), students receive an invitation by email to evaluate online all courses for which they partook in the examination. Since the first semester ends in January, this means that course evaluations for the first semester take place during the second semester, whereas course evaluations for the second semester take place over the summer break. If students do not respond within a number of weeks of receiving the evaluation email, they are sent a reminder email.

Each course evaluation consists of around 10 questions, the wording of which can differ slightly by faculty as well as across different degree programs within the same faculty. Table 1 gives an overview of the typical set of evaluation questions, covering teaching

Table 1 Evaluation questions

1	The teacher makes clear what knowledge and skills I should acquire to pass this course
2	The examination matches the proposed aims of the course (i.e. matches the the knowledge and skills the teacher states I should acquire)
3	The teaching method (e.g. lectures, assignments, usage of online learning environment) has helped me prepare for the course examination
4	The teaching method (i.e. lectures, tutorials, assignments, etc., taken together) stimulated me to participate actively
5	The study materials (slides, online learning environment, etc.) helped me study the course material
6	The program of study / learning track of material taught during contact hours was sufficiently clear and coherent
7	The teacher made efforts to make the course interesting
8	The teacher communicates clearly about practical matters and course organization
9	The teacher provided opportunities to assess my progress during the course (e.g. by welcoming questions, giving assignments or midterm exams, providing an online discussion forum, ..)
10	I am satisfied with the quality of teaching in this course
11	I was present regularly during the contact hours of this course (lectures, tutorials, ..)

Note: Each question is scored on a scale of 1 (worst score) to 6 (best score)

Table 2 Number of unique observations

Dimension	Observations
Students	28,240
Courses	3329
Teachers	1781
Evaluation questions	160
Faculties	14
Student-course-teacher	350,486
Student-course-question	3,230,321
Student-course-teacher-question	3,473,374

Note: The dimension question reflects the 160 different evaluation questions

style (“The teaching method stimulated me to participate actively”), course content (“The teacher makes clear what knowledge and skills I should acquire to pass this course”, “The learning track of material taught during contact hours was sufficiently coherent”), as well as course organization (“The teacher communicates clearly about practical matters and course organization”) and examination (“The examination matches the proposed aims of the course”). Also typically included are a broader statement about teaching quality (“I am satisfied with the quality of teaching in this course”) and an indication of effort put forward by the student (“I was present regularly during the contact hours of this course”).

The dataset covers all evaluated courses⁷ in the academic year 2010/2011, for 14 faculties divided into three broad faculty groups: Engineering, Science and Technology; Bio-Medical Sciences; and Humanities and Social Sciences.⁸ Observations are at the student-course-teacher-question level, and students, teachers, courses, and faculties are anonymized. The unique features of this dataset are its wide scope, and the fact that students who did not respond to the evaluation questions for a course they were enrolled in (as defined by exam participation) are also included.

Summary Statistics

Tables 2 and 3 show summary statistics for this dataset. Table 2 indicates the number of unique observations at various levels: in total, we observe 28,240 students in 3329 courses taught by 1781 teachers. Some students and teachers are observed in multiple faculties, and some courses are taught by multiple teachers.

Table 3 shows means and standard deviations for observed characteristics of courses, students and teachers.

The average course has an evaluation response rate of 45 %, confirming that non-response is quite common and highlighting the possibility of selection bias. The highest score that can be obtained for each evaluation question is 6, the lowest 1: however, this

⁷ Not all courses are evaluated in the academic year 2010/2011—however, the large majority of courses are covered.

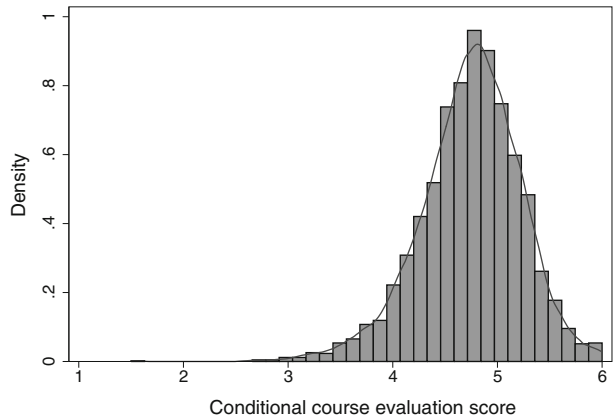
⁸ Engineering, Science and Technology includes the faculties of Science, Engineering Science, and Bio-science Engineering; Bio-Medical Sciences includes the faculties of Medicine, Pharmaceutical Science, and Kinesiology and Rehabilitation Sciences; and Humanities and Social Sciences includes the faculties of Theology and Religious Studies, Law, Economics and Business, Social Sciences, Arts, Psychology and Educational Sciences, and the Institute of Philosophy.

Table 3 Summary statistics

	Mean	Std dev
Course characteristics		
Response rate	45.0 %	19.4 %
Evaluation score	4.74	0.48
Grade (out of 20)	13.02	1.92
Pass rate	86.2 %	16.5 %
Course size	68.84	96.16
Percentage first semester	45.9%	49.8 %
Student characteristics		
Response rate	35.7 %	39.2 %
Item completion rate	95.1 %	15.5 %
Grade	11.82	3.22
Percentage female	55.4 %	49.7 %
Number of evaluated courses taken	8.12	3.35
Teacher characteristic		
Number of evaluated courses taught	2.67	2.00

Notes: Mean and standard deviation reported. The evaluation score lies between 1 (worst score) and 6 (best score), and the grade lies between 1 and 20, where a 10 or higher represents a pass. The response rate reflects whether a student has responded to a course evaluation by answering at least one question; the item completion rate reflects what percentage of evaluation questions are answered conditional on response for at least one question for a given course

Fig. 1 Course evaluation score distribution



Note: Line is a kernel density estimate. Course evaluation score observed conditional on response.

lowest score is seldom given as the average evaluation across courses is 4.74 with a standard deviation of 0.48. Figure 1 shows the distribution of evaluation scores at the course level.

Course grades are out of 20, where a grade of 10 or higher is a pass: the average course has a grade of 13.0 with a standard deviation of 1.92, and the average pass rate is 86 %. The grade we observe is the final one: this includes any passing grades resulting from retake exams that took place in the same academic year. The average course size is 69 students and slightly less than half of all evaluated courses are taught in the first semester.

At the student level, the response rate is some 36 %. Once a student has answered one evaluation question for a course, however, they almost always complete the full questionnaire: this is called the item completion rate in Table 3.

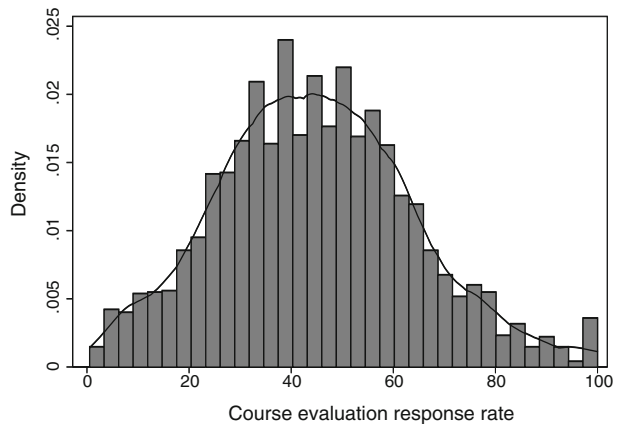
The average grade a student obtains is 11.8, with a standard deviation of 3.22. The average grade at the student level is lower than at the course level, reflecting that smaller courses typically have higher average grades. Around 55 % of all students are female and a student on average takes some 8 evaluated courses in the academic year.⁹ Furthermore, teachers on average teach 2.7 evaluated courses during the academic year.¹⁰

Finally, in addition to the low average response of 45 %, there is a large amount of variation in response rates across courses, as reflected by the standard deviation of 19 % points. Figure 2 shows the entire distribution of response rates across courses: this large variation implies that selection bias due to non-response may invalidate the common practice of comparing evaluation results across courses. The next section therefore outlines a selection model for course evaluations.

A Selection Model for SETs

The sample selection problem, first outlined by Gronau (1974) and Heckman (1974), arises whenever the outcome of interest is only observed for some subpopulation that is non-randomly selected. In the statistics literature on missing data, this is known as the problem of not-at-random missing data or non-ignorable non-response (Allison 2001; Rubin and Little 1987): see Berg (2005) for an overview of non-response bias. For SETs this particular problem arises when we observe evaluations only for a subset of students who are not randomly selected on observables such as grade, gender, and course size; and/or conditional on these observables, indicating selection on unobservable factors such as intrinsic motivation or opinions on the anonymity of SETs. Here we outline a Heckman selection model which includes selection along both observable and unobservable dimensions.

Fig. 2 Course evaluation response rate distribution



Note: Line is a kernel density estimate.

⁹ In reality, full-time students take 10 courses per year, reflecting that not all courses are evaluated in every year.

¹⁰ Note that the true number of taught courses will be somewhat higher since not all courses were evaluated in this academic year.

Response Equation

A student decides whether to respond to the evaluation based on her net utility, $Y_1^* \in (-\infty, +\infty)$, derived from response. This utility is determined by a vector of covariates, X_1 , and their coefficients, β_1 , as well as an additional term ε_1 :

$$Y_1^* = X_1\beta_1 + \varepsilon_1 \quad (1)$$

If $Y_1^* \geq 0$ the student responds and if $Y_1^* < 0$ the student does not respond to the course evaluation.

However, we do not observe Y_1^* , but only an indicator variable for whether the student decided to fill in the course evaluation. That is, we observe a variable Y_1 , which takes on the value 1 if the student evaluated the course and 0 otherwise:

$$\begin{aligned} Y_1 &= I(Y_1^* \geq 0) \\ &= I(X_1\beta_1 + \varepsilon_1 \geq 0) \end{aligned} \quad (2)$$

where $I(\cdot)$ denotes the indicator function, X_1 is observed by the econometrician, and ε_1 remains an unobserved error term that is assumed to be normally distributed with mean zero. This is the *response equation*. Coefficients in the response equation, β_1 , can be estimated from the full sample of all students that do and do not participate, i.e. $Y_1 = 1$ as well as $Y_1 = 0$, and their observable characteristics, X_1 .

What observed factors should we expect to be contained in X_1 , driving SET response? The existing SET literature points out that both high-achieving students (as measured by the (expected) grade or cumulative GPA) and, in most settings, female students are more likely to respond to course evaluations (Adams and Umbach 2012; Avery et al. 2006; Dey 1997; Kherfi 2011; Marsh 2007; Spooen and Van Loon 2012; Porter and Umbach 2006; Porter and Whitcomb 2005). In the case of achievement, this may be linked to salience arguments from a broader survey literature: if students do not wish to be reminded of a course in which they obtained a low grade, they may be less likely to respond to SETs (Adams and Umbach 2012). Furthermore, survey fatigue has been shown to decrease course evaluation response, where fatigue is the result of being presented with many surveys around the same time and/or overly long questionnaires (Adams and Umbach 2012; Groves et al. 2004; Spooen and Van Loon 2012; Porter et al. 2004). In addition, students who are less intensively involved in education, as evidenced by a light course-load, also tend to respond less (Kherfi 2011).

Whereas some of the factors driving SET response may be observed, not all need be (these being contained in ε_1). For example, students' ideals of study behavior (e.g. reflecting socialization and their intrinsic motivation for education) likely also play a role (Wolbring and Treischl 2015).¹¹ In particular, students with high intrinsic motivation may feel more invested in the educational process, increasing their response rate. Finally, students may be less likely to respond if they are concerned about anonymity of the evaluation (Dommeyer et al. 2004): whether a particular student has such concerns is usually unobserved as well.

¹¹ Part of this socialization effect may be captured by student gender, but other, unobserved, dimensions related to upbringing and student peer groups likely remain.

Evaluation Equation

A student’s evaluation of a course is a continuous variable $Y_2^* \in (-\infty, +\infty)$ that depends on a set of covariates, X_2 , and their coefficients, β_2 , as well as an additional term ε_2 :

$$Y_2^* = X_2\beta_2 + \varepsilon_2 \tag{3}$$

However, we do not observe Y_2^* but ordered integers $Y_2 \in \{1, 2, 3, 4, 5, 6\}$ instead,¹² which we will denote by v_h for $h = 1, \dots, 6$. To write the observed model in terms of the underlying latent model in Eq. (3), define $\kappa_0, \kappa_1, \dots, \kappa_5, \kappa_6$ with $\kappa_0 = -\infty$ and $\kappa_6 = +\infty$ and $\{\kappa_1, \dots, \kappa_5\}$ such that:

$$\forall h \in \{1, \dots, 6\} : \Pr(Y_2 = v_h) = \Pr(\kappa_{h-1} \leq Y_2^* < \kappa_h)$$

Given κ_h for $h = 0, \dots, 6$, we can then write Y_2 in terms of Y_2^* :

$$\begin{aligned} Y_2 &= \sum_{h=1}^6 v_h I(\kappa_{h-1} \leq Y_2^* < \kappa_h) \\ &= \sum_{h=1}^6 v_h I(\kappa_{h-1} \leq X_2\beta_2 + \varepsilon_2 < \kappa_h) \\ &\equiv V(Y_2^*) = V(X_2\beta_2 + \varepsilon_2) \end{aligned} \tag{4}$$

where v_h, κ_h , and X_2 are observed by the econometrician but ε_2 remains an unobserved error term that is assumed to be normally distributed with mean zero. The *evaluation equation* is notationally summarized by the non-linear expression $V(X_2\beta_2 + \varepsilon_2)$.

The evaluation scores can be observed only for the subgroup of students that decided to respond to the evaluation, i.e. Y_2 is only observed if $Y_1 = 1$. Therefore, the evaluation equation used to estimate β_2 is:

$$Y_2 | (Y_1 = 1) = V(X_2\beta_2 + \varepsilon_2) | \varepsilon_1 \geq -X_1\beta_1$$

Which factors influence students’ perception of teaching quality? There is a large literature studying the observable determinants (X_2) of SETs conditional on response. This literature has found several observed factors which correlate with conditional evaluation scores (see Marsh 2007 for an overview): prime among these is students’ achievement level. High-performing students, as evidenced by their grades, tend to rate courses more favorably. There are three prime explanations for this, which are not mutually exclusive. For one, students’ impression of teaching quality is expected to increase when students learn more, and grades are an indicator of learning. A second explanation is focused on grading leniency: if students receive lower grades for a given amount of learning, they rate the course less favorably (Marks 2000; McPherson and Jewell 2007), consistent with cognitive dissonance theories (Maurer 2006). Lastly, such achievement levels may reflect pre-existing differences between students (Marsh 2007). Course size, student and instructor gender, instructor rank, course level, and student personality also appear to affect observed course evaluations, although Marsh (2007) describes the evidence for these observed factors as mixed, and highlights that the found effects are typically small.

Since the found effects for observables can only explain a small part of the variation in evaluation scores (Spooen et al. 2013), one may also suspect that unobservables

¹² Reflecting the evaluation scores from 1 through 6 which are given in our dataset, see “Data description”.

(contained in ε_2) affect evaluation scores. For example, students’ intrinsic motivation could impact evaluation scores, since more motivated students are more likely to perceive their educational experience as rewarding and therefore rate educational quality more highly. This intrinsic motivation could also be linked to students’ prior subject interest, which has been measured in some institutional settings and indeed found to positively impact SETs conditional on response (Marsh 2007; Marsh and Dunkin 1997).

Selection Bias and the Selection Model

The average observed, or conditional, evaluation score, $E[Y_2|Y_1 = 1]$, is biased if it differs from its unconditional counterpart, that is, the mean evaluation score if all students had participated, $E[Y_2]$. For example, if students that participate in the course evaluation evaluate courses more generously for some reason, the average observed evaluation score will be upward biased. Based on our review of the literature, there are indeed likely to be factors which influence both the response and evaluation equation. Among observables, the student’s grade has previously been found to positively impact both: this may then lead to positive selection bias. Similarly, students’ unobserved intrinsic motivation could increase both response and evaluation scores, again leading to positive bias. If this is indeed the case, SET scores would not be representative of all students’ opinion on teaching quality. Furthermore, comparing the evaluation scores for two courses with identical teaching quality but different response rates would reveal a higher evaluation score for the course with the lower response rate. Lastly, estimated coefficients $\hat{\beta}_2$ may be biased.

This *selection bias* in evaluation scores due to non-response can further be decomposed into a bias from selection on observables and on unobservables:

$$\begin{aligned}
 \underbrace{E[Y_2|Y_1 = 1] - E[Y_2]}_{\text{total bias}} &= \underbrace{V(E[X_2\beta_2|\varepsilon_1 \geq -X_1\beta_1]) - V(E[X_2\beta_2])}_{\text{bias from observables}} \\
 &+ \underbrace{V(E[\varepsilon_2|\varepsilon_1 \geq -X_1\beta_1])}_{\text{bias from unobservables}} \tag{5}
 \end{aligned}$$

Equation (5) shows that bias from selection on observables exists if $E[X_2\beta_2|\varepsilon_1 \geq -X_1\beta_1] \neq E[X_2\beta_2]$ because the regressors in X_2 are a subset of X_1 . For example, if higher course grades predict higher probabilities of SET response as well as higher SET scores, $E[X_2\beta_2|\varepsilon_1 \geq -X_1\beta_1] > E[X_2\beta_2]$ and the observed evaluation scores are upward biased. Similarly, the final term in Eq. (5) shows that there is selection on unobservables if $E[\varepsilon_2|\varepsilon_1 \geq -X_1\beta_1] \neq E[\varepsilon_2] = 0$, i.e. if ε_1 and ε_2 are correlated. This bias occurs, for example, if more intrinsically motivated students are both more likely to respond and award higher scores.

However, the last two terms in Eq. (5), $V(E[X_2\beta_2])$ and $V(E[\varepsilon_2|\varepsilon_1 \geq -X_1\beta_1])$, are not observed by the econometrician. To quantify the total selection bias and its components, we therefore need a *selection model*. For strictly continuous outcome variables, simple two-step selection model estimators have been developed (Heckman 1978, 1979), but for ordinal responses, as in Eq. (4), accounting for sample selection is complicated by the fact that a non-linear model must be used to fit the data.¹³ Maximum likelihood (ML)

¹³ In particular, two-step procedures analogous to the Heckman (Heckman 1978, 1979) method are only approximate and no appropriate distribution results for the estimators are available. Hence, inference based on such procedures may lead to wrong conclusions (Heckman 1978; Van de Ven and Van Praag 1981; Wooldridge 2002).

techniques or two-stage method of moments are therefore needed (Miranda and Rabe-Hesketh 2006). In particular, we use Luca and Perotti's (2011) maximum likelihood procedure for implementing estimation. We identify this selection model by using an instrument, i.e. a variable that predicts SET response but not the course evaluation score. In other words, the instrument must be contained in X_1 but not in X_2 . All other observables may be included in both the response and evaluation equations.

From this selection model we obtain consistent estimates of β_2 . This means we can compare the estimated selection-corrected effects of observable characteristics (reflected by β_2) to their uncorrected counterparts: sizable differences would be problematic for a large related literature, surveyed in Marsh (2007), which analyzes the effects of observable student, course, and teacher characteristics on SETs without correcting for selection.

Furthermore, given that the regressors in X_2 are a subset of those in X_1 and are therefore observed for both responders as well as non-responders, the population wide average evaluation score can be predicted, $E[\hat{Y}_2]$. Consequently, the total bias on the left hand side of Eq. (5) can be estimated as the difference between the average observed score for responders only and the population wide average predicted score, $E[Y_2|Y_1 = 1] - E[\hat{Y}_2]$. The selection bias from observables, which is the first component on the right-hand side of Eq. (5), can then be obtained as the difference between the predicted average score conditional on response and the population wide average predicted score, $E[\hat{Y}_2|Y_1 = 1] - E[\hat{Y}_2]$. Lastly, the selection bias from unobservables is the difference between the previous two terms, $E[(Y_2 - \hat{Y}_2)|Y_1 = 1]$, which is the second component on the right-hand side of Eq. (5).

Besides quantifying the selection bias and its components in Eq. (5), the selection model estimated in “Assessing selection bias”, below, reports two additional statistics that are informative about the importance of selection bias. Firstly, a log likelihood ratio test statistic is reported, which compares the log likelihood of the full selection model with the sum of the log likelihoods for the evaluation and response equations estimated separately. A large log likelihood test statistic implies that the null hypothesis of no selection bias is rejected. Secondly, an estimate of the correlation between ε_1 and ε_2 (called ρ in Table 4 from “Estimating the selection model”) is reported. If this estimate is positive (negative) and significant, evaluation scores are upward (downward) biased because of positive (negative) selection on unobservables. Note, however, that the size of the bias in the average evaluation score ultimately depends not only on the size of this correlation but also on the amount of truncation due to non-response.

Assessing Selection Bias

Estimating the Selection Model

We now turn to estimating the selection model outlined in the previous section: as an instrument, we use the semester in which the course is taught. In exploratory results reported in a working paper version of this paper (Goos and Salomons 2014), we found that response for first-semester courses is significantly higher than for second-semester courses. This is probably the result of students being more likely to be on holiday during the evaluations of the second semester, or, in the case of last-year students, having graduated and left university. However, there is no reason to think that students' opinion on

Table 4 Selection-corrected and uncorrected models

	A. Selection-corrected model		B. Uncorrected model
	Response eq. (1)	Evaluation eq. (2)	Evaluation eq. (3)
Grade (stdized)	0.171*** (0.009)	0.163*** (0.012)	0.155*** (0.008)
Course passed	0.110*** (0.018)	0.029** (0.015)	0.025* (0.014)
Course size (stdized)	-0.033* (0.018)	-0.055*** (0.012)	-0.053*** (0.012)
Female student	0.218*** (0.017)	0.023 (0.016)	0.011 (0.012)
Number of evaluated courses taken by student (stdized)	0.103*** (0.010)	-0.010 (0.010)	-0.016* (0.008)
Number of evaluated courses taught by teacher (stdized)	0.013*** (0.005)	0.010** (0.005)	0.009* (0.005)
Course taught in first semester	0.346*** (0.024)	–	–
Observations	3,473,374		1,295,940
χ^2 statistic	2068.3***		2706.7***
ρ	0.078***		–
Likelihood Ratio statistic	12.92***		–

Notes: Panel A presents estimates of a Heckman selection model with a probit response equation and an ordered probit evaluation equation. Instrument: semester in which the course was taught. Panel B presents estimates of an ordered probit evaluation equation uncorrected for selection bias, i.e. it models the evaluation equation conditional on response. All independent variables, except for dummy variables, are standardized to have a zero mean and unit standard deviation. Two-way clustered standard errors, by student and by teacher, are reported in parentheses

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

courses depends on the semester in which they are taught, making the semester in which courses are taught a valid instrument.

Table 4 reports the estimates of this selection model: we follow Wooldridge (2002) by including all regressors that are in X_2 in X_1 also. Results are shown for our preferred specification, where a full set of observables is included: the student's grade, whether the student passed the course,¹⁴ the student's gender, the course size, and the number of evaluated courses taken by the student¹⁵ and taught by the teacher,¹⁶ respectively. However, our working paper version shows that results are very similar when only the most

¹⁴ The specification implies that the grade has a linear effect on both response rates and evaluation scores, with the exception of a discrete jump at the passing grade. We performed statistical tests to examine whether the effect of the grade is truly linear on either side of the passing rate and could not reject this hypothesis.

¹⁵ On the one hand, this can be seen as a proxy for survey fatigue, as in Adams and Umbach (2012); but it also measures the student's intensity of involvement in education, e.g. as a part-time or contract student versus as a full-time year-round student (Kherfi 2011). Its predicted effect on response is therefore ambiguous.

¹⁶ This could be interpreted as a measure of the extent to which a staff member's job is geared towards teaching as opposed to research.

important observable characteristic, the student's grade, is used. Panel A in Table 4 reports both the response and evaluation equation estimates.

From the response equation estimates in the first column of Table 4, it can be seen that the semester variable is a strong instrument: it highly significantly predicts SET response. Furthermore, the model's likelihood ratio test statistic shows that the null hypothesis of no selection is firmly rejected. This means there is statistically significant selection bias in the evaluation score as a result of non-response.

Table 4's panel A constitutes the first estimate of a Heckman selection model for SETs in the literature, and we can use its coefficient estimates for the evaluation equation to assess the validity of a large related body of work studying the observable determinants of SETs conditional on response (e.g. as surveyed in Marsh 2007 and Spooen et al. 2013). After all, selection bias implies that the estimates of the effects of observables on the evaluation score conditional on response may be biased, meaning they do not necessarily uncover the true unconditional effects. Therefore, panel B in Table 4 additionally presents the uncorrected ordered probit coefficients for our evaluation equation. There are only small differences between the corrected (column 2) and uncorrected (column 3) coefficients: this means that when one is interested in the observable determinants of SETs, correcting for selection bias does not appear to be essential, validating the literature's approach and findings.

However, this does not preclude the average SET score being biased: indeed, results from Table 4 suggest bias in the average score due to both observed and unobserved characteristics. Selection on observables results from the covariates reported in Table 4. In particular, the student's grade has a statistically significantly positive effect on both SET response and SET scores, and its effect is the largest of all covariates. The separate effect of passing the course on both response and evaluation is also positive and statistically significant. In contrast, the effects for other variables are typically smaller and not always statistically significant. Firstly, female students evaluate courses more often but do not give significantly higher scores. Also, larger courses elicit lower response and lower scores, although the effects are small. Furthermore, teachers who teach more evaluated courses have both slightly higher response rates and slightly higher scores for their course evaluations. Lastly, students who take more evaluated courses are more likely to participate but do not give significantly different scores. These found effects and their relative sizes are consistent with the literature, discussed in the Introduction and "A selection model for SETs".

The finding that, for all statistically significant estimates, the effects in the response and evaluation equations are identically signed implies that selection bias due to observable characteristics is positive. In other words, students who participate in the evaluation are more likely to have higher values for the observable characteristics (such as higher grades) that also positively affect the evaluation score. The same applies to the observable characteristics of courses with higher response rates and of teachers with higher response rates. This finding is consistent with the expectation of positive bias provided in Kherfi (2011) and Spooen and Van Loon (2012), as well as with Wolbring and Treischl (2015)'s finding of positive bias due to observables in a paper-based SET setting.

Selection on unobservables also leads to positive bias in the evaluation score: this is evidenced by the positive and statistically significant estimate of the correlation coefficient between the errors in the selection and evaluation equations, indicated by ρ in Table 4. In other words, students who participate in the evaluation have unobserved characteristics which increase their satisfaction with teaching, and/or courses and teachers with higher response rates have unobserved characteristics which also increase evaluation scores.

Table 5 Estimated selection bias

Total bias	0.1332
Bias from observables	0.0492
Bias from unobservables	0.0840

Note: Based on the Heckman selection model estimated in Table 4, panel A

Quantifying Selection Bias in SET Scores

The evidence in Table 4 shows that the direction of the bias in average SET scores is positive, both due to observed and unobserved characteristics. However, even more important than its direction is the size of the bias: is it large enough to invalidate interpreting the average course evaluation score as a measure of teaching quality in any one course, or to distort comparisons of evaluation scores across courses? We address these questions in the next two paragraphs.

Consequences for measuring teaching quality within courses To quantify the amount of total selection bias as well as the respective contributions from selection on observables and on unobservables, we use Eq. (5); results are reported in Table 5.

As already deduced qualitatively in “Estimating the selection model”, the average evaluation score is indeed upward biased. The magnitude of the total bias is found to be 0.1332. This corresponds to an economically non-negligible bias of around 28 % ($=0.1332/0.4834 \times 100\%$) of a standard deviation of the evaluation score across courses. Indeed, as a result of selection bias, the average course has an evaluation score that is higher by the same amount as having about a one standard deviation higher average grade (see Goos and Salomons 2014).

Moreover, Table 5 decomposes the total bias into the contributions from observables and unobservables as in Eq. (5). Here, unobservables are found to account for the majority of the total selection bias: specifically, 63 % ($0.084/0.1332 \times 100\%$). The relatively small upward selection bias in online SETs due to observables such as the grade is consistent with Wolbring and Treischl (2015)’s finding for paper-based SETs.

As a specification check, we have alternatively specified our model more flexibly by including interaction terms between all independent variables in the response and evaluation equations. Our results are not significantly different: in this case, we obtain an average total bias of 0.139 of which 0.050 is due to observed characteristics.

In sum, we find that course evaluations are upward biased as a result of non-response, such that the average evaluation score would be lower if all students participated in the evaluation. As such, teaching quality is not accurately measured in any one course. Furthermore, the comparison of evaluation scores across courses is also likely to be affected, since different courses have different response rates: we investigate this next.

Consequences for comparing teaching quality across courses One way to correct for selection bias in each course is to assess the relationship between the SET response rate and the amount of selection bias. This can be done by using the fact that the selection bias will, by definition, be zero at a response rate of 100 %. Since the total average selection bias of 0.1332 holds at the average response rate of 45 %, we find the bias for other response rates by connecting this point to a bias of zero at a response rate of 100 %, assuming the bias decreases linearly in the response rate.¹⁷ This exercise highlights that

¹⁷ By estimating course-level models and using an alternative instrument (the number of evaluated courses the student takes—results are reported in the Appendix of the working paper version of this paper), we have

selection bias is still around 20 % ($=0.1/0.4835 \times 100$ %) of a standard deviation of the evaluation score at a response rate of around 60 %. A 60 % response rate corresponds to the 79th percentile in the response rate distribution across courses, implying that some amount of bias remains even at relatively high response rates compared to the average of 45 %.

We can then use this analysis to correct the evaluation scores for all courses based on their response rate. For example, we find that courses in the bottom quartile of the response distribution on average have a 0.13-point lower SET score than courses in the top quartile of the response distribution, but after correcting for selection bias, this difference almost doubles to 0.25 points. Similarly, courses in the bottom quartile of the response distribution would on average move down 111 places in the SET-based ranking, whereas courses in the top quartile would on average move up 119 places in that ranking. Indeed, there are only 26 courses whose ranking is exactly the same before and after correcting for selection bias. For comparison, in the extreme case that evaluation scores are assigned completely at random, we would expect some 1100 rank changes for each course.¹⁸

The overall pattern of changes in the ranking of teaching quality based on SET scores is represented in Fig. 3. This figure plots how much the course rank changes, in absolute terms, for courses ordered by their observed SET score. In particular, it can be seen that the absolute change in the ranking is highest for courses in the middle of the SET score distribution, and lowest for courses at the extremes of the distribution. This is to be expected, since there is much more density around the mean than near the extremes of the distribution, making the ranking in the middle more sensitive to bias adjustment. Therefore, our results suggest that using SET scores for comparing teaching quality across courses is less problematic for the lowest- and highest-rated courses, as compared to the middle of the course evaluation distribution. Indeed, this provides an explanation for the finding in Jacob and Lefgren (2008), that SET-based rankings of teaching quality correspond quite well with the top and bottom of the teaching quality distribution according to objective measures of learning, but do a much worse job in the middle of the distribution.

Taken together, the results presented in this section suggest that the SET score in the average course does not accurately reflect the opinions of all students on teaching quality: indeed, students' impression of teaching quality is somewhat overestimated in this institution. Furthermore, without estimating a Heckman selection model, the majority of the bias goes undetected since it is mostly driven by unobserved characteristics. On the other hand, correcting for selection is not found to be very important when one is interested in effects of observables on SET scores. Lastly, selection bias affects the ranking of courses, although this problem is less pronounced at the extremes of the course evaluation distribution. The next section discusses the external validity of these findings as well as several limitations of our study.

Footnote 17 continued

tried to test this assumption: while there is some evidence that the relationship is somewhat convex, such that amount of bias decreases faster at low response rates, the found relationship from this alternative analysis is almost linear.

¹⁸ To find this counterfactual, we have simulated the distribution of observed course evaluation scores 1000 times, by randomly drawing from a normal distribution with a mean of 4.74 and standard deviation of 0.48 and calculating the evaluation rank of each course in each of these simulations. For each course and simulation, we then calculate the absolute rank difference between the actually observed rank and the one produced by the random assignment of course evaluation scores. Lastly, we average the absolute rank differences for each course across the 1000 trials.

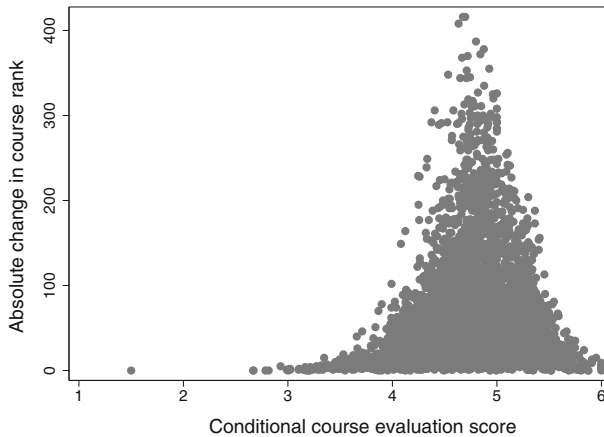


Fig. 3 Changes in SET course rank by conditional course evaluation score

External Validity and Limitations

An important limitation of this study is that it uses data from a single university: although this is quite typical in the literature on SETs,¹⁹ considering the specific institutional environment is important from the perspective of external validity.

Firstly, there may be some observable characteristics of students, teachers, and courses which matter for both SET response and evaluation scores but are not observed in our dataset. In this case, the effect of such observables may be included in our unobservable effect. Examples are the gender of the instructor or the level of the course, which we do not observe but may impact both response rates and evaluation scores. However, in surveying the literature, Marsh (2007) and Spooren et al. (2013) conclude that studies typically find only small roles for any particular set of observed characteristics in explaining variation in SETs. Indeed, across many studies, students grades are typically found to have the most robust effect, a finding we confirm, since variation in student grades is responsible for the large majority of bias due to observed characteristics. Moreover, in a non-SET setting, by comparing attrition patterns in longitudinal surveys of college students, Nielsen et al. (1978) also find that only part of students' non-response decisions can be predicted based on their set of observed characteristics. However, more research is needed to directly examine whether our findings extend to settings with a wider set of observables.

Secondly, we can compare the characteristics of the institution we examine with others that have been studied in the literature: our external validity increases to the extent that our setting is typical. There are several dimensions along which our institution does seem comparable to others. For example, average evaluation scores are typically found to be relatively high compared to the mean of the (Likert) scale (Nulty 2008; Wolbring 2012): our institution is no exception. Additionally, much variation in response rates across courses is also common (e.g. see Dommeyer et al. 2004), which implies that our institution is not atypical in this respect, and, more importantly, that any selection bias constitutes a problem for comparing SETs across courses in these other institutional settings as well.

¹⁹ With the notable exception of work on websites such as ratemyprofessor.com: e.g. see Felton et al. (2004, 2008); Timmerman (2008).

Furthermore, we include a wide range of different faculties and types of degree programs, eliminating any skewing of our results from focusing on a narrowly defined field of study or course type. Lastly, the average response rate at the institution we consider is at the higher end of the typical online SET response range of 30–50 % reported in other studies (e.g. see Adams and Umbach 2012; Layne et al. 1999; Nulty 2008; Spooren et al. 2013): this suggests that, if anything, more selection bias may be found in institutions with response rates at the lower end of this typical range.

However, our institution is perhaps less typical in other respects. For one, the timing of evaluations is atypical: these take place only after grades are known, rather than just before the end of term (Dommeyer et al. 2004).²⁰ This may lead us to overestimate the effect of the grade on evaluation scores, although *expected* grades have also been shown to impact evaluation scores (Marsh and Dunkin 1997). Previous work has shown that evaluations stabilize early on in the course (Costin 1968; Feldman 1979; Kohlan 1973), but it would still be interesting to compare the effect of grades on SETs in relationship with the timing of the survey to examine these effects more directly. Another institutional characteristic to consider is that SETs are web-based, and as such our results cannot be extended directly to paper-based SETs. As discussed in the Introduction, it is quite likely that selection mechanisms differ between online and paper-based evaluations, given that the latter are administered in a face-to-face setting with only a single opportunity to respond rather than online by logging in with a personal student ID over a longer period of time. However, all else equal, we expect selection bias to be less of a problem in paper-based SETs since average response rates are significantly higher. In any case, studying web-based SETs may not be a strong drawback since most institutions have adopted this mode of evaluation (Becker et al. 2011). Finally, the high item completion rate at this institution (reported in Table 3) is quite striking: this suggests there is little evidence of survey fatigue along this margin. In contrast, studies in different settings stress the importance of survey fatigue (both in terms of survey response and for item completion) in determining response rates (Adams and Umbach 2012). This means that we study selection mechanisms for survey response only, whereas in other settings, low item completion may be an additional issue.

Further limitations relate to our chosen methodology. In particular, empirically identifying Heckman selection models is challenging for a number of reasons. Firstly, without exclusion restrictions, identification is tenuous:²¹ this is why we have chosen the instrumental variable approach. Yet finding valid and strong instruments is no easy task, and what is valid in one institutional setting does not necessarily transfer to others. Furthermore, Heckman selection models are sensitive to misspecification (Pigini 2014): to deal with this, we have performed extensive robustness checks, reported in our working paper version. Lastly, parametric selection models such as ours rely on distributional assumptions about the error terms (in particular, joint normality) which may be violated, leading to inconsistent estimates (Arabmazar and Schmidt 1982; Manski 1989; Vella 1998). Although some work suggests that such violations may not lead to severe bias in practice (Van der Klaauw and Koning 2003), pursuing an alternative semi- or non-parametric approach may be a worthwhile extension of our work.

²⁰ However, online evaluations typically have long windows of opportunity during which students can respond, increasing the chance that these overlap with grade announcements in other settings also.

²¹ Strictly speaking, an instrument is not required since identification can come solely from distributional assumptions. However, as is well known, this is empirically problematic and therefore not recommended (e.g. see Keane 1992; Meng and Schmidt 1985; Puhani 2000).

Lastly, our model specification does not allow for any heterogeneity in the selection model across different courses. This implies, in particular, that we study the selection mechanisms only to the extent that they are common across courses, such as the possibility that more intrinsically motivated students are more likely to respond. This limitation particularly affects our counterfactual ranking analysis in “Quantifying selection bias in SET scores”. A more nuanced approach would be to estimate selection models at the course level, allowing selection mechanisms to differ across courses. However, such an exercise runs into practical problems, both because valid course-level instruments have to be found and because the number of observations in each course is often not sufficient for estimating selection models. Although we have estimated course-level models to fully account for such heterogeneity, and found similar results,²² our preferred instrument does not bind at the course level, making our setting less than ideally suited for examining this.

In sum, we believe our findings highlight that more research is needed into selection bias in SETs which may threaten the representativeness of such evaluations in any one course and hamper comparisons of teaching quality across courses with different response rates. By reporting the effect size (i.e. selection bias relative to the standard deviation of the evaluation score), we ensure that future studies can directly compare their findings to ours, and in doing so potentially reveal the importance of different institutional settings. Based on the limitations of our analyses, complementary follow-up work could explore the selectivity issue in other institutional settings, use a broader set of observables, pursue semi- or non-parametric models, as well as study any heterogeneity in selection mechanisms across different modes of evaluation or different courses.

Conclusions

Course evaluations often suffer from low response rates, and more so if the evaluation is online: we have argued how this may distort results, limiting the interpretation of course evaluations as a measure of teaching quality in any given course as well as rendering comparisons across courses, teachers, departments, and institutes problematic when the response rate varies. This is a first-order problem not yet sufficiently considered in the literature, despite SET scores commonly being used by departments for the purposes of awarding faculty teaching prizes and making promotion decisions, by students in course selection decisions, and for institutional rankings and accountability.

For a large European university, we indeed find that evaluations somewhat misrepresent student opinion about teaching quality—in particular, we find positive selection bias on average, indicating that the true evaluation score is lower. Furthermore, the SET-based ranking of courses is not accurate because the response rate varies widely across courses. As such, the findings in this study caution against taking SETs at face value, especially when average response rates are low or vary widely across courses. Indeed, in terms of SET-based course rankings, a cautious approach may be to create comparison bins where courses with similar response rates are grouped together. Furthermore, because course

²² As a robustness check, we have estimated these course-level selection models, using the less data-demanding Limited Information Maximum Likelihood (“two-step”) estimator and the number of evaluated courses taken by students as an instrument (since the semester instrument does not vary at the course level). Although the amount of bias at the course level was often imprecisely estimated, the average selection bias in courses where the model could be identified was found to be 0.1932—this is similar to the results, reported in Table 5, when we estimate selection models for all courses simultaneously using the semester as an instrument.

rankings at the extremes of the course evaluation distribution are found to be least affected, we suggest only identifying the 10 % best and worst performers rather than making judgments about teaching quality from differences across the middle of the observed SET score distribution.

Although previous work has suggested adjusting evaluation scores for student grades to avoid incentives for grade inflation, we find that this adjustment does not eliminate selection bias. This is because the majority of the bias is attributable to selection on other, unobserved, characteristics. To the extent that our observed characteristics are typical, then, this also implies that adjusting the course evaluation for selection bias is no easy task, and therefore unlikely to be adopted by university administrators on a large scale. On the other hand, we do not find evidence to suggest that the existing literature's practice of studying the effects of observable determinants of SET scores without adjusting for selection leads to biased results. This implies that estimating selection models may not be strictly necessary when one is interested in the effects of observables on SET scores.

The overall implication of our findings is that institutions should devote serious efforts to increasing SET response rates. From the literature, several possible strategies emerge. For example, Dommeyer et al. (2004) find that SET response rates are significantly higher when a very small grade incentive is offered. Also, as is common in some universities already, students may only receive their grade in a timely fashion if an evaluation is submitted (Dommeyer et al. 2004). Another avenue may be to perform online evaluations in class, similar to the way paper-based evaluations are administered, since this has been shown to increase response rates (Kordts-Freudinger 2013). Furthermore, email reminders and assurances about student anonymity may be important for online SETs especially (Dommeyer et al. 2002; Layne et al. 1999)—see also Cook et al. (2000) for a meta-analysis of strategies for increasing response rates in web-based surveys more generally. Additionally, we suggest that a scheme of random sampling might help improve the representativeness of the evaluation. In such a scheme, respondents are randomly selected rather than relying on students' own response decisions. Crucial to this scheme is that students sign a contract at the start of their studies which obliges them to fill in a course evaluation whenever they are randomly selected to participate in one. This scheme has the advantage of not increasing the average student's time spent on filling in questionnaires, thereby reducing survey fatigue, while still generating evaluation results which can be used to gauge students' opinion on teaching quality within and across courses (provided, of course, that standard errors are also reported).

Acknowledgments This paper has benefited from helpful comments by Annelies Govaerts and Veerle Hulpiau; participants of the Leuven Economics of Education Research (LEER) workshop; participants in seminars at Utrecht School of Economics, the Tinbergen Institute in Amsterdam, and Maastricht University; as well as by two anonymous reviewers.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Adams, M. J., & Umbach, P. D. (2012). Nonresponse and online student evaluations of teaching: Understanding the influence of salience, fatigue, and academic environments. *Research in Higher Education*, 53(5), 576–591.

- Allison, P. D. (2001). *Missing data. Quantitative applications in the social sciences*. USA: SAGE Publications.
- Alter, M., & Reback, R. (2014). True for your school? How changing reputations alter demand for selective U.S. colleges. *Educational Evaluation and Policy Analysis*, 36(3), 346–370.
- Angrist, J. D., & Guryan, J. (2008). Does teacher testing raise teacher quality? Evidence from state certification requirements. *Economics of Education Review*, 27(5), 483–503.
- Arabmazar, A., & Schmidt, P. (1982). An investigation of the robustness of the tobit estimator to non-normality. *Econometrica*, 50(4), 1055–1063.
- Avery, R. J., Bryant, W. K., Mathios, A., Kang, H., & Bell, D. (2006). Electronic course evaluations: Does an online delivery system influence student evaluations? *The Journal of Economic Education*, 37(1), 21–37.
- Becker, W. E., Bosshardt, W., & Watts, M. (2011). Revisiting how departments of economics evaluate teaching. Technical report, Working paper presented at the annual meetings of the American Economic Association.
- Becker, W. E., & Watts, M. (1999). How departments of economics evaluate teaching. *American Economic Review*, 89(2), 344–349.
- 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.
- Berg, N. (2005). Non-response bias. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (Vol. 2, pp. 865–873). London: Elsevier.
- Berk, R. A. (2005). Survey of 12 strategies to measure teaching effectiveness. *International Journal of Teaching and Learning in Higher Education*, 17, 48–62.
- Braga, M., Paccagnella, M., & Pellizzari, M. (2014). Evaluating students' evaluations of professors. *Economics of Education Review*, 41, 71–88.
- Carrell, S. E., & West, J. E. (2010). Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy*, 118(3), 409–432.
- Centra, J. A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness*. San Francisco: Jossey-Bass.
- Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, 104(9), 2593–2632.
- Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, 104(9), 2633–2679.
- 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.
- Coleman, J., & McKeachie, W. (1981). Effects of instructor/course evaluations on student course selection. *Journal of Educational Psychology*, 73, 224–226.
- Cook, C., Heath, F., & Thompson, R. L. (2000). A meta-analysis of response rates in web- or internet-based surveys. *Educational and Psychological Measurement*, 60(6), 821–836.
- Costin, F. (1968). A graduate course in the teaching of psychology: Description and evaluation. *Journal of Teacher Education*, 19, 425–432.
- Cunha, J. M. & Miller, T. (2014). Measuring value-added in higher education: Possibilities and limitations in the use of administrative data. *Economics of Education Review*, 42(C):64–77.
- Dey, E. (1997). Working with low survey response rates: The efficacy of weighting adjustments. *Research in Higher Education*, 38(2), 215–227.
- Dommeyer, C. J., Baum, P., Chapman, K. S., & Hanna, R. W. (2002). College students' attitudes towards methods of collecting teaching evaluations: In-class vs online. *Journal of Education for Business*, 78(1), 11–15.
- Dommeyer, C. J., Baum, P., Hanna, R. W., & Chapman, K. S. (2004). Gathering faculty teaching evaluations by in-class and online surveys: Their effects on response rates and evaluations. *Assessment and Evaluation in Higher Education*, 29(5), 611–623.
- Ewing, A. M. (2012). Estimating the impact of relative expected grade on student evaluations of teachers. *Economics of Education Review*, 31(1), 141–154.
- Feldman, K. (1979). The significance of circumstances for college students' ratings of their teachers and courses. *Research in Higher Education*, 10(2), 149–172.
- Felton, J., Koper, P. T., Mitchell, J., & Stinson, M. (2008). Attractiveness, easiness and other issues: Student evaluations of professors on RateMyProfessors.com. *Assessment and Evaluation in Higher Education*, 33(1), 45–61.
- Felton, J., Mitchell, J., & Stinson, M. (2004). Web-Based student evaluations of professors: The relations between perceived quality, easiness and sexiness. *Assessment and Evaluation in Higher Education*, 29(1), 91–108.

- Galbraith, C. S., Merrill, G. B., & Kline, D. M. (2012). Are student evaluations of teaching effectiveness valid for measuring student learning outcomes in business related classes? A neural network and bayesian analyses. *Research in Higher Education*, 53(3), 353–374.
- Goos, M., & Salomons, A. (2014). *Measuring teaching quality in higher education: Assessing the problem of selection bias in course evaluations*. TKI Discussion Paper 14–16, Utrecht School of Economics.
- Greenwald, A. G., & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist*, 52, 1209–1217.
- Gronau, R. (1974). Wage comparisons—a selectivity bias. *Journal of Political Economy*, 82(6), 1119–1143.
- Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. Hoboken, NJ: Wiley.
- Hanushek, E. A., Kain, J. F., O'Brien, D. M., & Rivkin, S. G. (2005). The market for teacher quality. NBER Working Papers 11154, National Bureau of Economic Research, Inc.
- Hanushek, E. A. and Rivkin, S. G. (2006). *Teacher Quality*, volume 2 of *Handbook of the economics of education* (pp. 1051–1078), chapter 18. Elsevier.
- Heckman, J. J. (1974). Shadow prices, market wages, and labor supply. *Econometrica*, 42(4), 679–694.
- Heckman, J. J. (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica*, 46(4), 931–959.
- Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1), 153–161.
- Ho, D. E., & Shapiro, T. H. (2008). Evaluating course evaluations: An empirical analysis of a Quasi-experiment at the stanford law school, 2000–2007. *Journal of Legal Education*, 58(388).
- Hoffmann, F., & Oreopoulos, P. (2009). Professor qualities and student achievement. *The Review of Economics and Statistics*, 91(1), 83–92.
- Imberman, S. A. & Lovenheim, M. (2013). Does the market value value-added? Evidence from housing prices after a public release of school and teacher value-added. CESifo Working Paper Series 4105, CESifo Group Munich.
- Isely, P., & Singh, H. (2005). Do higher grades lead to favorable student evaluations? *The Journal of Economic Education*, 36(1), 29–42.
- Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics*, 26(1), 101–136.
- Johnson, R. (2000). The authority of the student evaluation questionnaire. *Teaching in Higher Education*, 5(4), 419–434.
- Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review*, 27(6), 615–631.
- Keane, M. P. (1992). A note on identification in the multinomial probit model. *Journal of Business and Economic Statistics*, 10(2), 193–200.
- Kherfi, S. (2011). Whose opinion is it anyway? Determinants of participation in student evaluation of teaching. *The Journal of Economic Education*, 42(1), 19–30.
- Kogan, L. R., Schoenfeld-Tacher, R., & Hellyer, P. W. (2010). Student evaluations of teaching: Perceptions of faculty based on gender, position, and rank. *Teaching in Higher Education*, 15(6), 623–636.
- Kohlman, R. G. (1973). A comparison of faculty evaluation early and late in the course. *Journal of Higher Education*, 44(8), 587–595.
- Kordts-Freudinger, R., & Geithner, E. (2013). When mode does not matter: Evaluation in class versus out of class. *Educational Research and Evaluation*, 19(7), 605–614.
- Krautmann, A. C., & Sander, W. (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), 417–428.
- Layne, B. H., DeCristoforo, J. R., & McGinty, D. (1999). Electronic versus traditional student ratings of instruction. *Research in Higher Education*, 40(2), 221–232.
- Leventhal, L., Abrami, P., & Perry, R. (1976). Do teacher rating forms reveal as much about students as about teachers? *Journal of Educational Psychology*, 68, 441–445.
- Leventhal, L., Abrami, P., Perry, R., & Breen, L. (1975). Section selection in multi-section courses: Implications for the validation and use of teacher rating forms. *Educational and psychological measurement*, 35, 885–895.
- Liegle, J. O., & McDonald, D. S. (2004). Lessons learnt from online vs paper-based computer information students' evaluation system. *Information Systems Education Journal*, 3(37), 3–14.
- Liu, Y. (2006). A comparison study of online versus traditional student evaluation of instruction. *International Journal of Instructional Technology and Distance Learning*, 4, 15–29.
- Luca, G. D., & Perotti, V. (2011). Estimation of ordered response models with sample selection. *Stata Journal*, 11(2), 213–239.

- MacNell, L., Driscoll, A., & Hunt, A. (2015). Whats in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, 40(4), 291–303.
- Manski, C. (1989). Anatomy of the selection problem. *Journal of Human Resources*, 24(3), 343–360.
- Marks, R. B. (2000). Determinants of student evaluations of global measures of instructor and course value. *Journal of Marketing Education*, 22(2), 108–119.
- Marsh, H. W. (2007). *Students' evaluations of university teaching: A multidimensional perspective*. chapter 9. The scholarship of teaching and learning in higher education: an evidence-based perspective (pp. 319–384). Springer.
- Marsh, H. W. & Dunkin, M. J. (1997). Students' evaluations of university teaching: A multidimensional perspective. chapter 9. *Effective teaching in higher education: research and practice* (pp. 241–320). Agathon.
- Maurer, T. W. (2006). Cognitive dissonance or revenge? Student grades and course evaluations. *Teaching of Psychology*, 33(3), 176–179.
- McPherson, M. A. (2006). Determinants of how students evaluate teachers. *The Journal of Economic Education*, 37(1), 3–20.
- McPherson, M. A., & Jewell, R. (2007). Leveling the playing field: Should student evaluation scores be adjusted? *Social Science Quarterly*, 88(3), 868–881.
- 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.
- Meng, C.-L., & Schmidt, P. (1985). On the cost of partial observability in the bivariate probit model. *International Economic Review*, 26(1), 71–85.
- Miranda, A., & Rabe-Hesketh, S. (2006). Maximum likelihood estimation of endogenous switching and sample selection models for binary, ordinal, and count variables. *Stata Journal*, 6(3), 285–308.
- Morley, D. D. (2012). Claims about the reliability of student evaluations of instruction: The ecological fallacy rides again. *Studies in Educational Evaluation*, 38(1), 15–20.
- Nielsen, H., Moos, R., & Lee, E. (1978). Response bias in follow-up studies of college students. *Research in Higher Education*, 9(2), 97–113.
- Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: What can be done? *Assessment and Evaluation in Higher Education*, 33(3), 301–314.
- Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, 26(3), 237–257.
- Pigini, C. (2014). Bivariate non-normality in the sample selection model. *Journal of Econometric Methods*, 4(1), 123–144.
- Pope, A. (2014). The effect of teacher ratings on teacher performance. Technical report, mimeo University of Chicago.
- Porter, S., & Umbach, P. (2006). Student survey response rates across institutions: Why do they vary? *Research in Higher Education*, 47(2), 229–247.
- Porter, S. R., E., W. M., & Weitzer, W. H. (2004). Multiple surveys of students and survey fatigue (pp. 63–74). *Overcoming survey research problems: vol. 121. New directions for institutional research*. Jossey-Bass.
- Porter, S. R., & Whitcomb, M. E. (2005). Non-response in student surveys: The role of demographics, engagement and personality. *Research in Higher Education*, 46(2), 127–152.
- Puhani, P. (2000). The Heckman correction for sample selection and its critique. *Journal of Economic Surveys*, 14(1), 53–68.
- Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73(2), 417–458.
- Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, 94(2), 247–252.
- Rubin, D. B., & Little, R. J. A. (1987). *Statistical analysis with missing data*. New York: Wiley.
- Shih, T.-H., & Fan, X. (2009). Comparing response rates in e-mail and paper surveys: A meta-analysis. *Educational Research Review*, 4(1), 26–40.
- Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research*, 83(4), 598–642.
- Spooren, P. & Van Loon, F. (2012). Who participates (Not)? A non-response analysis on students' evaluations of teaching. *Procedia—social and behavioral sciences*, vol. 69 (pp. 990–996). International Conference on Education and Educational Psychology (ICEEPSY 2012).
- Timmerman, T. (2008). On the validity of RateMyProfessors.com. *Journal of Education for Business*, 84(1), 55–61.
- Van de Ven, W. P. M. M., & Van Praag, B. M. S. (1981). The demand for deductibles in private health insurance: A probit model with sample selection. *Journal of Econometrics*, 17(2), 229–252.

- Van der Klaauw, B., & Koning, R. H. (2003). Testing the normality assumption in the sample selection model with an application to travel demand. *Journal of Business and Economic Statistics*, *21*(1), 31–42.
- Vella, F. (1998). Estimating models with sample selection bias: A survey. *Journal of Human Resources*, *33*(1), 127–169.
- Weinberg, B. A., Hashimoto, M., & Fleisher, B. M. (2009). Evaluating teaching in higher education. *The Journal of Economic Education*, *40*(3), 227–261.
- Wilhelm, W. B. (2004). The relative influence of published teaching evaluations and other instructor attributes on course choice. *Journal of Marketing Education*, *26*(1), 17–30.
- Wolbring, T. (2012). Class attendance and students' evaluations of teaching: Do no-shows bias course ratings and rankings? *Evaluation Review*, *36*(1), 72–96.
- Wolbring, T. & Treischl, E. (2015). Selection bias in students' evaluation of teaching. *Research in Higher Education*, 1–21.
- Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. MIT Press.
- Zabaleta, F. (2007). The use and misuse of student evaluations of teaching. *Teaching in Higher Education*, *12*(1), 55–76.