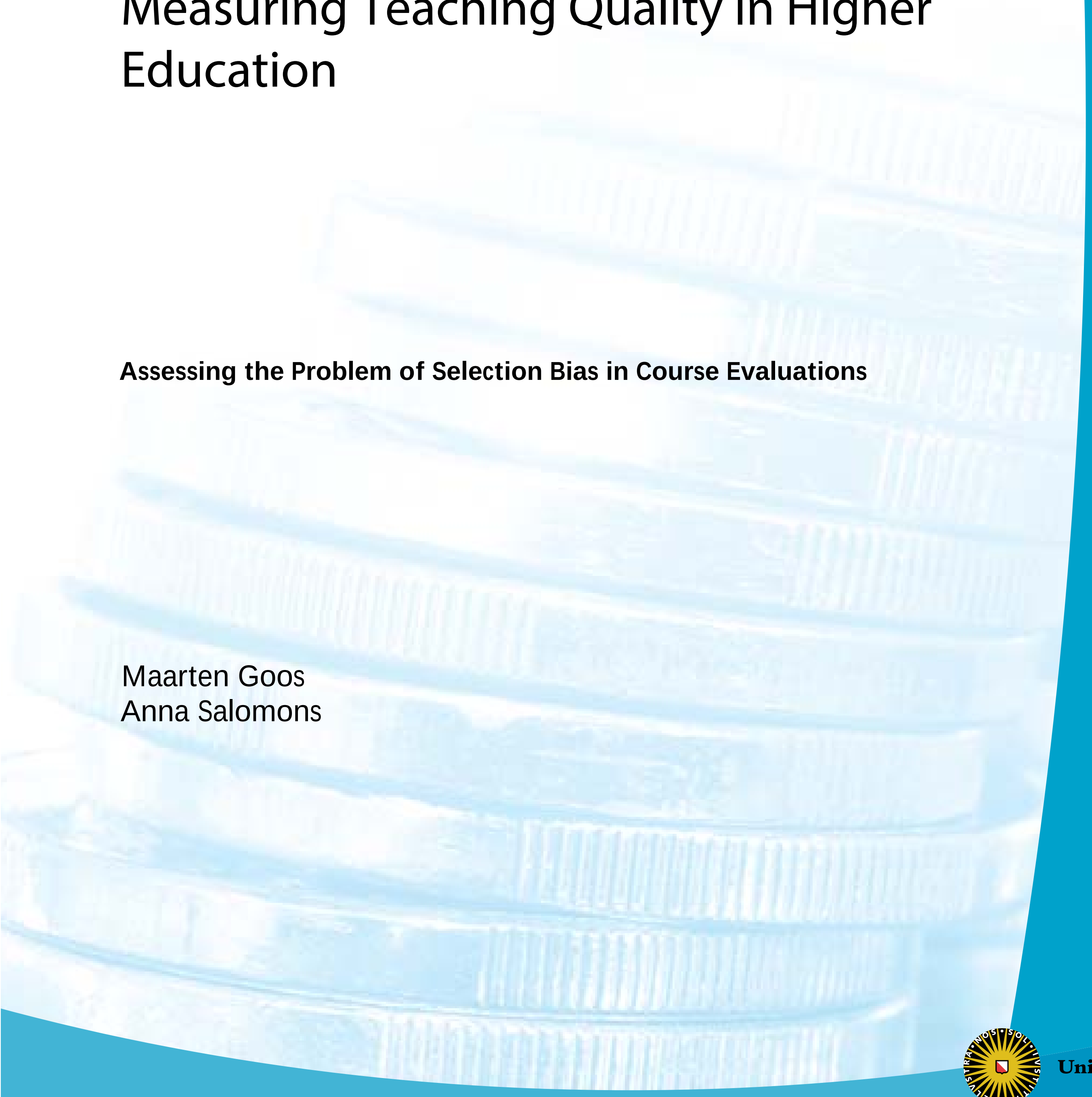
Tjalling C. Koopmans Research Institute

Tjalling C. Koopmans

Discussion Paper Series nr: 14-16

Measuring Teaching Quality in Higher Education

Assessing the Problem of Selection Bias in Course Evaluations



Maarten Goos
Anna Salomons

Universiteit Utrecht

Utrecht School of Economics

Tjalling C. Koopmans Research Institute
Utrecht School of Economics
Utrecht University

Kriekenpitplein 21-22
3584 EC Utrecht
The Netherlands

telephone	+31 30 253 9800
fax	+31 30 253 7373
website	www.koopmansinstitute.uu.nl

The Tjalling C. Koopmans Institute is the research institute and research school of Utrecht School of Economics.
It was founded in 2003, and named after Professor Tjalling C. Koopmans, Dutch-born Nobel Prize laureate in economics of 1975.

In the discussion papers series the Koopmans Institute publishes results of ongoing research for early dissemination of research results, and to enhance discussion with colleagues.

Please send any comments and suggestions on the Koopmans institute, or this series to J.M.vanDort@uu.nl

ontwerp voorblad: WRIK Utrecht

How to reach the authors

Please direct all correspondence to the last author.

Maarten Goos
Center for Economic Studies
KU Leuven
Naamsestraat 69 - bus 3565
3000 Leuven
Belgium
E-mail: maarten.goos@kuleuven.be
Anna Salomons
Utrecht University
Utrecht School of Economics
Kriekenpitplein 21-22
3584 TC Utrecht
The Netherlands.
E-mail: a.salomons@uu.nl

This paper can be downloaded at: http://
www.uu.nl/rebo/economie/discussionpapers

Utrecht School of Economics
Tjalling C. Koopmans Research Institute
Discussion Paper Series 14-16

Measuring Teaching Quality in Higher Education: Assessing the Problem of Selection Bias in Course Evaluations

Maarten Goos[a]
Anna Salomons[b]

[a]Center for Economic Studies
KU Leuven

[b]Utrecht School of Economics
Utrecht University

December 2014

Abstract

Student evaluations of teaching are widely used to measure teaching quality and compare it across different courses, teachers, departments and institutions: as such, they are of increasing importance for teacher promotion decisions as well as student course selection. However, the response on course evaluations is rarely perfect, rendering such uses unwarranted if students who participate in the evaluation are not randomly selected: this paper is the first to investigate this issue. We quantify the direction and size of selection on both observable and unobservable characteristics for a large European university where course evaluation response rates differ across courses. Our results suggest course evaluations are upward biased, and that this bias mostly derives from selection on characteristics unlikely to be observed by the typical university administrator. Correcting for selection bias has sizable effects on both scores in any given course and the evaluation-based ranking of different courses.

Keywords: Educational economics, Student evaluations of teaching (SET), Education quality, Sample selection

JEL classification: I23, J24

1 Introduction

Education quality matters for learning outcomes (Hanushek, Kain, O'Brian & Rivkin 2005) as well as outcomes later in life (Chetty, Friedman & Rockoff 2014b), and an important aspect of education quality is teacher quality (Rockoff 2004; Nye, Konstantopoulous & Hedges 2004; Hanushek et al 2005). However, there is less agreement on the appropriate measurement of teacher quality, particularly for higher education where the availability of standardized testing to directly compare teaching performance is rare.[3] Yet public interest in this measurement is rising because college attainment has increased markedly in many countries, and since institutions of higher education to a large extent rely on scarce public resources.

Two broad measuring approaches are used for gauging teacher quality in higher education. Firstly, the certification of teachers: however, there is little evidence that certified teachers are more effective (Rivkin et al 2005; Kane, Rockoff & Staiger 2008; Angrist & Guryan 2008).[4] The second measure for teacher performance in higher education are course evaluations filled in by students. Usage of this second measure is by far the most widespread (Becker & Watts 1999, Becker, Bosshardt & Watts 2011).[5] Such course evaluations are used on a large scale to assess the quality of instruction at institutions for higher education, and also for comparing teacher performance across courses, departments and even universities (Becker et al 1999, 2011). The evaluations affect published institutional teaching rankings[6], and are used as input for promotion decisions.

However, the use of course evaluations as a measure of teaching quality has been criticized for a number of reasons. Firstly, it is often argued that the signal course evaluations provide

[3] At the primary or secondary level, economists often measure teacher quality 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, Friedman & Rockoff 2014a,b), although there is some evidence that parents do not respond to these (Imberman & Lovenheim 2014, Pope 2014), and use of these indicators is not (yet) widespread in any case. These measures are not applicable at the level of higher education, because tests are not standardized - an essential component for the measurement of teacher value added. See Cunha & Miller (2014) for an in-depth exploration of the possibilities and limitations of value-added measures in higher education.

[4] However, there is evidence to suggest that teachers who are more qualified in the field they are teaching are more effective than so-called adjunct teachers (e.g. see Carrell & West 2010).

[5] Other approaches such as peer evaluation also occur, but this is used only sporadically and generally given little weight in assessments of teaching quality, unlike course evaluations which are both highly used and highly weighted (Becker et al 1999, 2011).

[6] 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. There is evidence that such rankings affect student college applications (Alter & Reback 2014).

on teacher quality is contaminated by noise. Indeed, evaluation results tend to reflect (course or teacher) characteristics which may not be related to teaching quality (e.g. see Berk 2005; Isely & Singh 2005; McPherson, Jewell & Kim 2009), suggesting that the ability of students to assess the quality of teaching provided to them is limited.[7] Related to this is the argument that such noise in course evaluations provides teacher incentives for grade inflation, since there is much evidence to suggest that the average (expected) grade has a positive effect on course evaluations irrespective of learning outcomes (Krautmann & Sander 1999, McPherson 2006, Langbein 2008, Ewing 2012). 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 many observable course characteristics to filter out the noise, as suggested by Greenwald & Gillmore (1997), McPherson (2006) and Mcpherson et al (2009), among others.[8]

In this paper, we address a more fundamental concern arising from the measurement of teaching quality by means of course evaluations: possible selection bias resulting from non-response. In most institutions, students are not required to fill out course evaluations. As a result, response rates are rarely 100 percent, and often even below 50 percent, which raises the concern that the results are not representative of all students' opinions. 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, importantly, compare teaching quality across courses or teachers with different response rates. 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, or analyses of teacher- or course-

[7] 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 later learning outcomes). There is as yet no consensus about this, however (Clayson 2009). For instance, Weinberg, Fleisher & Hashimoto (2008), Carrell & West (2010) and Braga, Paccagnella & Pellizzari (2014) find that student evaluations are not correlated to learning, as measured by performance in follow-up courses. Furthermore, Morley (2012) finds that measures of interrater agreement are not always high. On the other hand, Hoffman and Oreopolous (2006) find that evaluations predict teacher quality as measured by drop-out rates, and Beleche, Fairris & Marks (2012) find a positive, albeit small, correlation between evaluations and subsequent student learning.

[8] Against the claim that course evaluations contain no signal with respect to teaching quality, one may argue that there is still value in knowing student utility (on which course evaluations inform, see Braga, Paccagnella & Pellizzari 2014) even if it is not directly informative about teaching quality.

level correlations between learning outcomes and course evaluations. Therefore, analyzing selection bias is a first-order concern. Lastly, this concern is exacerbated in online evaluations, which are increasingly used for their cost-effectiveness (Becker et al 1999, 2011) but typically have significantly lower response rates (Liegle & McDonald 2004, Avery, Bryant, Mathios, Kang & Bell 2006, Ho & Shapiro 2008, Shih & Fan 2009).

This first-order question has remained largely unaddressed in the literature, possibly because data on non-respondents is typically unavailable. Kherfi (2011) and Spooren & Van Loon (2012) study the determinants of course evaluation response, comparing the observable characteristics of respondents to those of non-respondents. Both find significant differences: for example, respondents on average have higher grades. Further, in an analysis of the determinants of higher course evaluation 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. This study goes beyond these approaches and findings in the literature, firstly, by explicitly quantifying the selection bias (in terms of both sign and size) using a selection model, and secondly, by analyzing selection on unobservables rather than only observables in an instrumental variables approach, and lastly, by using a much larger and richer dataset.[9]

Therefore, in this paper, we set out to make two contributions. Firstly, we investigate the size and direction of selection bias in course evaluations: for this, we use detailed information from all courses at a large European university in which education was evaluated in the academic year 2010/2011. Secondly, we contrast the findings for selection on unobservables to those for selection on observables. After all, if selection bias is primarily a matter of selection on student and teacher characteristics that most university administrators can observe, it can in principle be corrected for. Correcting for selection on unobservables is a decidedly more difficult exercise, which is unlikely to be adopted by institutions on a large scale.

This paper is structured as follows. In the next section, we succinctly outline the selection problem for course evaluations in a case that covers both selection on observables and on unobservables. The third section describes the dataset and provides summary statistics as well as exploratory

[9] Previous studies observe at most a few thousand students, whereas our study covers close to thirty thousand across all departments.

analysis. Results from estimated selection models are outlined and discussed in the fourth section. The final section concludes and offers some tentative policy implications based on our findings.

2 A selection model for course evaluations

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. For course evaluations this particular problem arises when we observe evaluations only for a subset of students which 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 ability or motivation. Here we outline a set-up which combines selection on observables and on unobservables.

2.1 Participation equation

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

$$Y_1^* = X_1\beta_1 + \varepsilon_1 \tag{1}$$

If $Y_1^* \geq 0$ the student participates and if $Y_1^* < 0$ the student does not participate in 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 or not. That is, we observe a variable Y_1, which takes on the value 1 if the student evaluated the course and 0 otherwise:

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

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 *participation equation*. Coefficients in the participation 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.

2.2 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 an integer $Y_2 \in \{1, 2, 3, 4, 5, 6\}$ instead, which we will denote by v_h for $h = 1, ..., 6$. To write the observed model in terms of the underlying latent model in equation (3), define $\kappa_0, \kappa_1, ..., \kappa_5, \kappa_6$ with $\kappa_0 = -\infty$ and $\kappa_6 = +\infty$ and $\{\kappa_1, ..., \kappa_5\}$ such that:

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

Given κ_h for $h = 0, ..., 6$, 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 nonlinear expression $V(X_2\beta_2 + \varepsilon_2)$.

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

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

2.3 Selection bias and the selection model

The average observed evaluation score, $E[Y_2|Y_1 = 1]$, is biased if it differs from the mean evaluation score if all students had participated, $E[Y_2] = V(E[X_2\beta_2])$. For example, if students that participate in the course evaluation are students that evaluate courses more generously for some reason, the average observed evaluation score will be upward biased. If this is the case, comparing the evaluation scores for two teachers of courses with different response rates would give a higher score to the teacher with the lower response rate, even if the teaching quality in both courses is identical.

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

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

What equation (5) shows is that a bias from selection on observables exist 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.[10] Assume, for example, that we observe a student's course grade and that higher course grades predict higher probabilities of participation in course evaluation as well as higher course evaluations. If this is the case, $E[X_2\beta_2|\varepsilon_1 \geq -X_1\beta_1] > E[X_2\beta_2]$ and the observed evaluation scores are upward biased. Similarly, the final term in equation (5) shows that there is selection on unobservables if $E[\varepsilon_2|\varepsilon_1 \geq -X_1\beta_1] \neq E[\varepsilon_2] = 0$ because ε_1 and ε_2 are correlated.

However, the last two terms in equation (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

[10] We follow Wooldridge (2002) by including all regressors that are in X_2 also in X_1.

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 equation (4), accounting for sample selection is complicated by the fact that a nonlinear model must be used to fit the data.[11] Maximum likelihood (ML) techniques or two-stage method of moments are therefore needed (Miranda & Rabe-Hesketh 2005). In particular, we use De Luca & Perotti's (2011) maximum likelihood procedure for implementing estimation. To identify this selection model, an instrument is needed that predicts participation in the course evaluation but not the course evaluation score.[12] 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 participation and evaluation equations.

From this selection model we obtain consistent estimates of β_2. Given that the regressors in X_2 are a subset of those in X_1 and are therefore observed for both participants as well as non-participants, the population wide average evaluation score can be predicted, $E[\widehat{Y}_2]$. Consequently, the total bias on the left hand side of equation (5) can be estimated as the difference between the average observed score from participants only and the population wide average predicted score, $E[Y_2|Y_1 = 1] - E[\widehat{Y}_2]$. The selection bias from observables, which is the first component on the right hand side of equation (5), can then be obtained as the difference between the predicted average score conditional on participation and the population wide average predicted score, $E[\widehat{Y}_2|Y_1 = 1] - E[\widehat{Y}_2]$. Lastly, the selection bias from unobservables is the difference between the previous two terms, $E[(Y_2 - \widehat{Y}_2)|Y_1 = 1]$, which is the second component on the right hand side of equation (5).

Besides quantifying the selection bias and its components in equation (5), the selection model estimated in Section 4, 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 participation equations estimated seperately. A large log likelihood test statistic implies that

[11] In particular, two-step procedures analogous to the 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; de Ven & van Praag 1981; Wooldridge 2002). However, in the appendix, we provide traditional two-step estimates assuming our outcome models are linear for comparison (Appendix Tables 5A-5C and 6A-6C): our results are robust to this.

[12] 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 Meng & Schmidt 1985, Keane 1992).

the null hypothesis of no selection bias is rejected. Secondly, an estimate of the correlation between ε_1 and ε_2 (called "Rho" in Table 5 from Section 4) is reported. If this estimate is positive (negative) and significant, evaluation scores are upwards (downwards) biased because of positive (negative) selection on unobservables.

3 Data description

3.1 Institutional background

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 they took part in. 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 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 courses[13] in the academic year 2010/2011, for 14 faculties divided into three broad departmental groups: Science, Engineering and Technology; Bio-Medical Sciences; and Humanities and Social Sciences. Observations are at the student-course-teacher-

[13] Not all courses are evaluated in the academic year 2010/2011- however, the large majority of courses are covered.

question level, and students, teachers, courses and faculties are anonymized. Other than its wide scope, the unique feature of the dataset is that students who did not respond to the course evaluation questions for a course they took are also included.

3.2 Summary statistics

Tables 2A and 2B show summary statistics for this dataset. Table 2A indicates the number of unique observations at various levels: in total, we observe 28,243 students in 3,329 courses taught by 1,781 teachers.[14] The dataset is very rich, covering all students and staff across 3 groups of faculties: Science, Engineering and Technology (5,802 students); Biomedical Sciences (7,446 students); and Humanities and Social Sciences (17,592 students).

Table 2B shows means and standard deviations for observable characteristics of courses, students and teachers. The average course has an evaluation response rate of 45 percent, with a standard deviation of 19 percentage points- this highlights the possibility of selection bias due to non-response. It can be seen that variation in the response rate is not the result of variation in average response across department groups: in each group, the average response rate is close to the overall average of 45 percent. However, within each group, there is a large amount of variation in response rates as reflected by the standard deviations of 17 to 20 percentage points. Figure 1 shows the distribution of response rates across courses: the large variation in response rates implies that any selection bias due to non-response would invalidate comparing evaluation results across courses.

The highest score that can be obtained for each evaluation question is 6, the lowest 1: however, this lowest score is rarely given as the evaluation for the average course is 4.7 with a standard deviation of 0.48. Figure 2 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.[15] Grades for courses taught in Humanities and Social Science faculties are lowest on average (12.8), and highest for courses taught in Bio-Medical Science faculties (13.4). The corresponding average pass rate is 86 percent for all courses, ranging

[14] Some students and teachers are observed in multiple faculties, and some courses are taught by multiple teachers.

[15] 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.

between an average of 88 percent for courses in Bio-Medical Sciences to 85 percent for courses in Humanities and Social Sciences.

The average course size is 69 students (with a standard deviation of 96), and this average varies between around 45 students in Science, Engineering and Technology to 77 students in Humanities and Social Sciences. Lastly, slightly less than half of all evaluated courses are taught in the first semester in all department groups.

At the student level, the response rate is some 36 percent. Once a student has answered one evaluation question for a course, however, they almost always complete the full questionnaire: this is called the intensive margin response rate in Table 2B. On average, students fully complete 95 percent of the evaluations they start.

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 percent of all students are female: this percentage is only 33 percent in the Science, Engineering and Technology department group, 60 percent in Humanities and Social Sciences, and 67 percent in the Bio-Medical Sciences. Lastly, a student on average takes some 8 evaluated courses in the academic year.[16]

Teachers on average teach 2.7 evaluated courses during the academic year: in Humanities and Social Sciences, this is slightly higher at 3 courses; in Science, Engineering and Technology it is 2.8 courses; and in Bio-Medical sciences, it is lowest at 2.4 courses.[17]

3.3 The determinants of participation

What factors affect participation in the course evaluation? This subsection examines this by providing exploratory evidence on the first equation of the selection model. We estimate a linear

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

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

probability model as the participation equation introduced in Section 2.1:

$$\begin{aligned} Y_{1ict} &= \beta_0 + \beta_1 grade_{ict} + \beta_2 pass_{ic} + \beta_3 female_i + \beta_4 nrcourses_i \\ &\quad + \beta_5 nrcourses_t + \beta_6 size_c + \beta_7 first_c + \varepsilon_{ict} \end{aligned} \tag{6}$$

where Y_{1ict} is a dummy for whether the student i answered at least one evaluation question for course c and teacher t. This model does not exploit any variation across evaluation questions since we found that the intensive margin response rate is close to 100 percent. Note that some courses have multiple teachers (which at times also assign different grades).

$Grade_{ict}$ is the grade student i obtained in course c taught by teacher t; $pass_{ic}$ is a dummy indicating that student i passed course c; $female_i$ is a dummy which takes on the value 1 when the student is female; $nrcourses_i$ is the number of evaluated courses a student took up in the academic year 2010/2011; $nrcourses_t$ is the number of evaluated courses a teacher taught in the academic year 2010/2011; $size_c$ is the course size; and $first_c$ is a dummy indicating that course c was taught in the first rather than second semester of the academic year. All variables, except for the dummies, have been standardized to have a zero mean and unit standard deviation.

The first column of Table 3 estimates equation (6): faculty, teacher, course and student dummies are respectively added to the specification in subsequent columns. The last columns present estimates where multiple sets of dummies are added simultaneously. Note that, depending on the set of dummies added, variation from one or more of the covariates in equation (6) will be fully absorbed.

The estimates reported in Table 3 show that grades are positively correlated with participation in the evaluation in many specifications. For example, in column 1, which does not include any fixed effects, a one standard deviation higher grade increases participation by 6.3 percentage points, all else equal. This result also applies to students within the same course: in columns 4 and 7, where course dummies are included in the specification, a one standard deviation increase in the grade increases the participation probability by 7.3 percentage points. However, this picture is less clear-cut when student fixed effects are introduced (as in the specifications of columns 5, 8, 10 and

11). Here, negative and statistically insignificant effects can be found, indicating that the same student does not respond more often to the course evaluation for courses where she obtained a higher grade, all else equal.

A more consistent pattern can be found for the effect of passing a course: the between-student variation suggests that students who pass a course are more likely to fill in the evaluation by a margin of 3.4 to 5.5 percentage points, and for the same student, passing a course also increases the probability of participation by 1.6 to 2.2 percentage points.

Female students are more likely to fill in the evaluation, also within the same faculty (column 2), for the same teacher (column 3), within courses (column 4) as well as combinations of these (columns 6, 7 and 9). The effect ranges between 6.3 to 7.7 percentage points, a sizable difference.

Furthermore, students who took more evaluated courses are more likely to participate in the evaluation. This could of course be explained mechanically, but the effect also holds within the same course (columns 4 and 7), suggesting that one standard deviation more evaluated courses taken increases the probability of participation in the evaluation of any one course by around 4.5 percentage points.

The number of evaluated courses taught by the teacher affects the response rate only modestly, increasing response by a little over 1 percentage point for each one standard deviation increase.

The effect of course size on the response rate is typically negative, but not very large: courses that are one standard deviation larger experience 0.5 to 2.1 percentage point lower response rates. The exception to this is the effect of course size for a given teacher: larger courses taught by the same teacher actually achieve slightly higher response rates (by 1.4 percentage points) compared to smaller courses taught by the same teacher (and this holds also within the same faculty). This indicates that teachers may have some influence over the response rate of courses.

Lastly, courses that are taught in the first semester obtain response rates that are a sizable 10.5 to 12.4 percentage points higher than those taught in the second semester: this finding is extremely robust across different model specifications.[18]

[18] Appendix Table 1 reports estimates for equation (6) separately for each of the three departmental groups: the results discussed in this section also apply within each of these groups.

3.4 The determinants of evaluation scores

Having examined what correlates with participation in the course evaluation, we now turn to the evaluation equation outlined in Section 2.2. In specific, what student and course characteristics correlate with higher course ratings, conditional on participation? To consider this, we estimate the following model using OLS:

$$\begin{aligned} Y_{2iqct}|\,(Y_{1iqct}=1) \quad &= \quad \gamma_0+\gamma_1 grade_{ict}+\gamma_2 pass_{ic}+\gamma_3 female_i+\gamma_4 nrcourses_i \\ &\quad\; +\gamma_5 nrcourses_t+\gamma_6 size_c+\gamma_7 first_c+u_{iqct} \end{aligned} \tag{7}$$

where Y_{2iqct} is the evaluation score given by student i to evaluation question q for course c and teacher t, which is only observed conditional on participation, i.e. conditional on $Y_{1iqct}=1$. The regressors are as before. We add fixed effects to subsequent specifications in the same way as for the participation equation but, unlike that equation, this model does exploit variation across evaluation questions since the same student can rate a course differently for different evaluation questions.

Table 4 provides estimates of equation (7). The evidence suggests that the observed evaluation score is positively influenced by the grade a student obtains. In the specification without fixed effects, a one standard deviation in grade leads to a sizable 0.14-point increase in the course evaluation score, which corresponds to 30 percent of a standard deviation in the observed evaluation score at the course level. This effect is even stronger in specifications that control for student fixed effects (columns 5, 8, 10 and 11), indicating that the same student rates the courses she obtained higher grades in more highly: around 0.20-point higher evaluation score for a one standard deviation higher grade. In addition to the grade, passing the course also increases the evaluation score in every specification, by some 0.05 to 0.10 points.

Female students typically give slightly higher observed evaluation scores, also within the same course, but the effect is not very large (0.02 to 0.05 points).

Students who take more evaluated courses typically give slightly lower evaluation scores: although often statistically significant, this effect is very small indeed (0.01 to 0.02 points for each one standard deviation increase in the number of courses). Similarly, teachers who teach more evaluated courses obtain slightly higher evaluation scores (0.01 to 0.02 points).

Larger courses have lower observed evaluation scores: a one standard deviation increase in course size decreases evaluation scores by up to 0.05 points. Lastly, the effect on the observed evaluated score of the semester the course is taught is small and typically insignificant.[19]

4 Assessing selection bias

4.1 Estimating selection models

Having explored the participation and evaluation equations separately, we now turn to estimating selection models, which, by estimating these equations jointly, allow determining the sign and size of selection bias in course evaluations. As an instrument, we use the semester in which the course is taught. As reported in the previous section, 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 courses depends on the semester in which they are taught, making the semester in which courses are taught a valid instrument.[20]

Table 5 reports estimates of these selection models, using three different specifications. The first column reports results when no observables are included in the equations. The second includes the grade as an observable, since this was found to be the most important determinant of the evaluation score in Section 3.4. The third and final column reports results from our preferred specification, where a full set of observables is included: the student's grade, whether the student passed the course, the student's gender, the course size, and the number of evaluated courses taken by the

[19] Appendix Table 2 reports equation (7) estimated separately for each departmental group, with very similar results.

[20] Alternatively, we have used the number of evaluated courses the student takes as an instrument. The reasoning underlying this instrument is as follows: when a student takes more evaluated courses, she is more likely to fill in an evaluation at some point during the academic year, increasing the probability of response. In the previous section, we found this to be the case even within courses. However, we do not expect students who take more evaluated courses to necessarily have a different opinion about any given course. The results from using this instrument are qualitatively identical (although quantitatively, slightly larger selection biases are found) to using the semester as an instrument, and are available in Appendix Tables 3 and 4. Also reported in these appendix tables are results when the semester and the number of courses are both used as instruments.

student and taught by the teacher, respectively. Both the participation and evaluation equation estimates are reported.

From the participation equation estimates in Table 5, it can be seen that the semester variable is a strong instrument: it highly significantly predicts participation in the course evaluation in all three specifications. Furthermore, the likelihood ratio test statistics show that for all three specifications of the selection model, the null hypothesis of no selection is firmly rejected. This means there is significant selection bias in the evaluation score as a result of non-response.

The significance of these selection models imply two things. Firstly, the estimates of the effects of observables on the evaluation score conditional on response are biased. This means that a selection model should be used to gauge the amount of selection bias due to observables. Secondly, there is significant selection bias due to unobservable characteristics. Here, we will first discuss these two sources of bias in turn, and subsequently quantify them.

Selection on observables results from the covariates reported in Table 5. The student's grade has a significant positive effect on both participation and on the evaluation score, and its effect is the largest of all covariates. The effect of passing the course on both participation and evaluation is also positive. The effects for other variables are statistically significant but smaller: female students evaluate courses more often and also give higher scores; larger courses elicit lower response and lower scores; teachers who teach more courses have both higher response rates and higher scores for their course evaluations; and lastly, students who take more courses are more likely to participate but give slightly lower scores (although this last effect is not significant at the 1 percent level). The finding that, for all but one covariate, the effects on participation and evaluation are identically signed implies that selection bias due to observable characteristics is positive. For example, the students who participate in the evaluation are more likely to have higher values for observable characteristics (such as higher grades) which also positively covary with the evaluation score. The same applies to the observable characteristics of courses and teachers with higher participation rates.

Selection on unobservables also leads to positive bias in the evaluation score: this is evidenced by the positive estimate of the correlation coefficient between the errors in the selection and evaluation equations, indicated by Rho in Table 5. In other words, students who participate in the evaluation have unobservable characteristics which increase their satisfaction with teaching, and courses and teachers with higher participation rates have unobservable characteristics which also increase evaluation scores.

4.2 Quantifying the bias

Besides signing the selection bias, the selection model also allows quantifying the amount of total selection bias as well as the respective contributions from selection on observables and on unobservables, as in equation (5): results are reported in Table 6. As before, the estimated selection bias is shown for the three different models (no covariates, grade as the only covariate, and the full set of covariates).

As already deduced qualitatively, the total evaluation score is indeed upward biased. The magnitude of the total bias is 0.2249 in the model without covariates, and decreases to 0.1297 once the grade is included in the specification. Since the grade accounts for most of the explanatory power among covariates, there is little difference between the estimated selection bias from the specification with only the grade as covariate and with the full set of covariates. Our preferred specification is the model which controls for all covariates, where the selection bias is found to be 0.1332. This corresponds to an economically sizable bias of around 28 percent (=0.1332/0.4834*100%) of a standard deviation of the evaluation score across courses.[21] As a result of selection bias, the average course, therefore, has an evaluation score that is higher by the same amount as having about a one standard deviation higher average grade.[22]

Moreover, Table 6 decomposes the total bias into the contributions from observables and unobservables as in equation (5): this is only relevant for the models in columns 2 and 3, where observables are included in the specification. Here, unobservables are found to account for the majority of the total selection bias: 65 percent (0.085/0.1297*100%) in the model which controls for the grade, and 63 percent (0.084/0.1332*100%) in the preferred model which controls for all covariates.

[21] The 0.48 standard deviation of the evaluation score across courses is reported in Table 2B.

[22] Based on the estimates reported in column 1 of Table 4, a one standard deviation increase in the grade increases the evaluation score by 0.14 points.

4.3 Consequences for measuring teaching quality

We have found that course evaluations are significantly 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. To quantify this, selection models would have to be estimated at the course level. However, this is not very feasible, both because valid course-level instruments would have to be found and because the number of observations in each course is often not sufficient for estimating selection models.[23]

Another way to correct for selection bias in each course is to assess the relationship between the response rate on the course evaluation and the amount of selection bias. This can be done by using the fact that the selection bias will, by definition, by zero at a response rate of 100 percent.

Figure 3 shows the results of this exercise. Since the total average selection bias holds at the average response rate of 45 percent, we draw a line connecting this point to a bias of zero at a response rate of 100 percent, assuming the bias decreases linearly in the response rate. This is done for each of the three models estimated in Table 5. For our preferred model, this figure shows that the selection bias is still around 20 percent (=0.1/0.4835*100%) of a standard deviation in the evaluation score at a response rate of some 60 percent. A 60 percent 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 response rate of 45 percent.

We can then use this simple analysis to correct the evaluation scores for all courses, based on their response rate. Figure 4 shows both the original and adjusted distributions of course evaluation scores, for the three different models. This illustrates the positive bias in course evaluation score, but

[23] 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 as 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 the paper, when we estimate selection models for all courses simultaneously.

also further informs on the comparison problems caused by imperfect response. For example, courses in the bottom quartile of the response rate distribution on average have a 0.13-point lower course evaluation than courses in the top quartile of the response rate distribution, but after correcting for selection bias, this difference almost doubles to 0.25 points. Similarly, courses in the bottom quartile of the response rate distribution would on average move down 111 places in the evaluation ranking (out of 3,329 courses), whereas courses in the top quartile would on average move up 119 places in that ranking. In other words, not only is the average score different, as evidenced by the distribution in Figure 4, the relative location of courses in this distribution has also changed. Economically, this means both the absolute observed evaluation score cannot be interpreted as truly reflecting student opinion on teaching quality, and the relative ranking of different courses (or their teachers) based on student evaluations is unwarranted.

5 Conclusions

Course evaluations often suffer from low response rates, more so if the evaluation is online: we have argued how this may distort results, limiting the interpretation of course evaluations in any given course as well as rendering comparisons across courses, teachers, faculties and universities problematic. This is a first-order problem not yet considered in the literature, which focuses on analyzing the correlates of evaluation scores while taken the representativeness of the scores themselves as given.

For a large European university, we indeed find that evaluations misrepresent student opinions about teaching quality- in particular, we find positive selection bias on average, indicating that the true evaluation score is lower. This bias is mostly attributable to selection on unobservables, although we do find a strong positive effect of the grade on the course evaluation. Even though the size of the bias would be expected to decrease in the response rate, conservative estimates suggest significant bias remains even at relatively high response rates. In sum, the findings in this study caution against taking student evaluations of courses at face value, especially when response rates are low or vary widely across courses.

We expect our covariates to be typical of what university administrators observe. This implies that, generally, adjusting the course evaluation for bias is no easy task. Although previous work has suggested adjusting evaluation scores for observables, we find that adjusting course evaluations for observable characteristics does not account for the majority of selection bias. Furthermore, in the presence of significant selection on unobservables, the unbiased effect of observables on the evaluation score can also only be identified by estimating a selection model.

The implication is of course that institutions should attempt to increase the response rate. For example, a scheme of random sampling might help improve the representativeness of the evaluation: students are randomly selected rather than relying on students' own participation 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. Alternatively, as is common in some universities already, the randomly selected student only sees their grades in a timely fashion if an evaluation is submitted. This scheme has the advantage of not increasing the average student's time spent on filling in questionnaires, while still generating evaluation results which can be compared across courses (provided, of course, that standard errors are also reported).

References

Avery, R. J., Bryant, W.K., Mathios, A., Kang, H., & D. Bell (2006). Electronic Course Evaluations: Does an Online Delivery System Influence Student Evaluations? *Journal of Economic Education,* 37(1): 21–38.

Alter, M. & R. Reback (2014). True for Your School? How Changing Reputations Alter Demand for Selective U.S. Colleges. Forthcoming *Educational Evaluation and Policy Analysis.*

Becker, W. E. & M. Watts (1999). How Departments of Economics Evaluate Teaching. *American Economic Review Papers and Proceedings,* 89(2): 344–349.

Becker, W. E., Bosshardt, W. & M. Watts (2011). Revisiting How Departments of Economics Evaluate Teaching. Working paper presented at the annual meetings of the American Economic Association.

Beleche, T., Fairris, D. & M. Marks (2012). Do Course Evaluations Truly Reflect Student Learning?

Evidence from an Objectively Graded Post-Test. *Economics of Education Review,* 31: 709–719.

Berk, R. A. (2005). Survey of 12 Strategies to Measure Teaching Effectiveness. *International Journal of Teaching and Learning in Higher Education,* 17(1): 48–62.

Braga, M., Paccagnella, M., & M. Pellizzari (2014). Evaluating Students' Evaluations of Professors. *Economics of Education Review,* 41: 71–88.

Carrell, S. E. & J. E. West (2010). Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors. *Journal of Political Economy,* 118(3): 409–432.

Chetty, R., Friedman, J. N. & J. E. Rockoff (2014a). Measuring the Impacts of Teachers Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review,* 104(9): 2593–2632.

Chetty, R., Friedman, J. N. & J. E. Rockoff (2014b). Measuring the Impacts of Teachers Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review,* 104(9): 2593–2632.

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.

Cook, C., Heath, F. & R. L. Thompson (2000). A Meta-Analysis of Response Rates in Web- or Internet-Based Surveys. *Educational and Psychological Measurement,* 60(6): 821–836 .

Cunha, J. M. & T. Miller (2014). Measuring Value-Added in Higher Education: Possibilities and Limitations in the Use of Administrative Data. *Economics of Education Review,* 42: 64–77.

De Luca, G., & V. Perotti (2011). Estimation of Ordered Response Models with Sample Selection. *Stata Journal* 11: 213–239.

De Ven, W. V., & B. Van Praag (1981). The Demand for Deductibles in Private Health Insurance: A Probit Model with Sample Selection. *Journal of Econometrics,* 17: 229–252.

Ewing, A. M. (2012). Estimating the Impact of Relative Expected Grade on Student Evaluations on Teachers. *Economics of Education Review,* 31: 141–154.

Isely, P. & H. Singh (2005). Do Higher Grades Lead to Favorable Student Evaluations?. *Journal of Economic Education,* 36(1): 29–42.

Greenwald, A. G. & G. M. Gillmore (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.

Liegle, J. O. & D. S. McDonald (2004). Lessons Learnt from Online vs Paper-Based Computer Information Students’ Evaluation System. *Information Systems Education Journal* 3(37).

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica,* 47(1): 153–161.

Heckman, J. J. (1978). Dummy Endogenous Variables in a Simultaneous Equation System. *Econometrica,* 46: 931–959.

Heckman, J. J. (1974). “Shadow Prices, Market Wages, and Labor Supply,” *Econometrica,* 42(4): 679–94.

Ho, D. E. & T. H. Shapiro (2008). Evaluating Course Evaluations: An Empirical Analysis of a Quasi-Experiment at the Stanford Law School, 2000-2007. 58 *Journal of Legal Education,* 388.

Hoffmann, F. & P. Oreopoulos (2009). Professor Qualities and Student Achievement. *The Review of Economics and Statistics*, 91(1): 83–92.

Imberman, S. A. & M. Lovenheim (2014). Does the Market Value Value-Added? Evidence from Housing Prices after Public Release of Teacher Value-Added“ NBER WP No. 19157

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. & W. Sander (1999). Grades and Student Evaluations of Teachers. *Economics of Education Review,* 18: 59–63.

Langbein, L. (2006). Management by Results: Student Evaluation of Faculty Teaching and the Mis-Measurement of Performance. *Economics of Education Review,* 27: 417–428.

McPherson, M. A. (2006). Determinants of How Students Evaluate Teachers. *Journal of Economic Education,* 37: 3–20.

McPherson, M. A., Jewell, R. T., & M. Kim (2009). What Determines Student Evaluation Scores? A Random Effects Analysis of Undergraduate Economics Classes. *Eastern Economic Journal,* 35: 37–51.

Miranda, A. & S. Rabe-Hesketh (2006). Maximum Likelihood estimation of Endogenous Switching and Sample Selection Models for Binary, Ordinal, and Count Variables. *The 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: 15–20.

Pope, A. (2014). The Effect of Teacher Ratings on Teacher Performance. mimeo University of Chicago http://home.uchicago.edu/~/npope/la_ny_paper.pdf.

Sartori, A. E. (2003). An Estimator for Some Binary-Outcome Selection Models Without Exclusion Restrictions. *Political Analysis,* 11: 111–138.

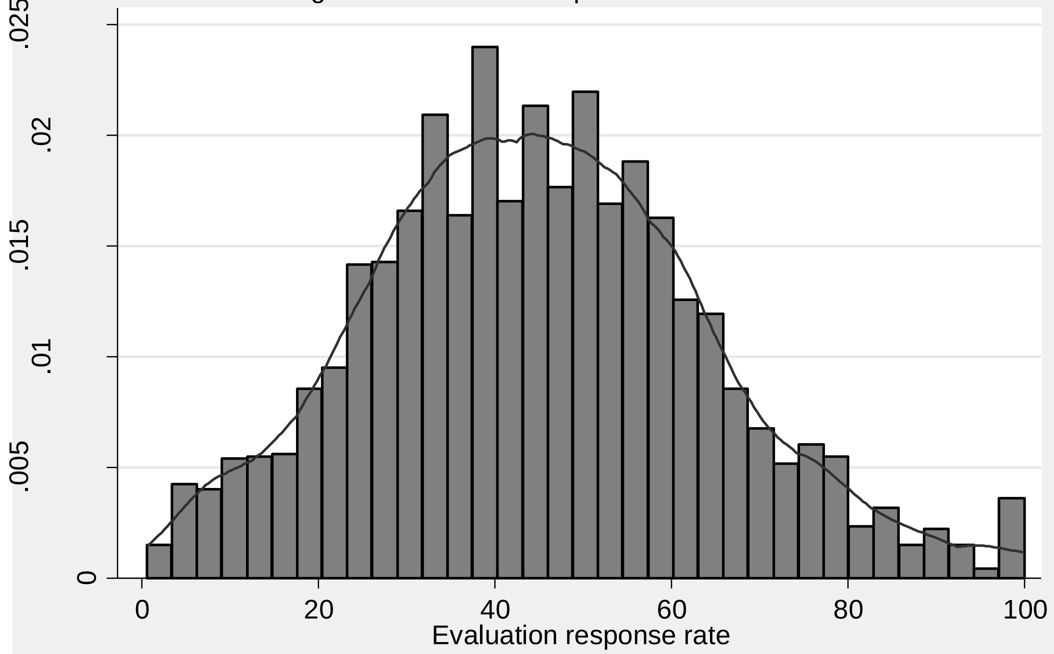
Shih, T. & X. Fan (2009). Comparing Response Rates in E-mail and Paper Surveys: A Meta-Analysis. *Educational Research Review,* 4(1): 26–40.

Weinberg, B., Hashimoto, M. & B. M. Fleisher (2009). Evaluating Teaching in Higher Education. *Journal of Economic Education,* 40(3): 227–261.

Wooldridge, J. M. (2002). Econometric Analysis of Cross Section and Panel Data. Cambridge, MA: MIT Press.

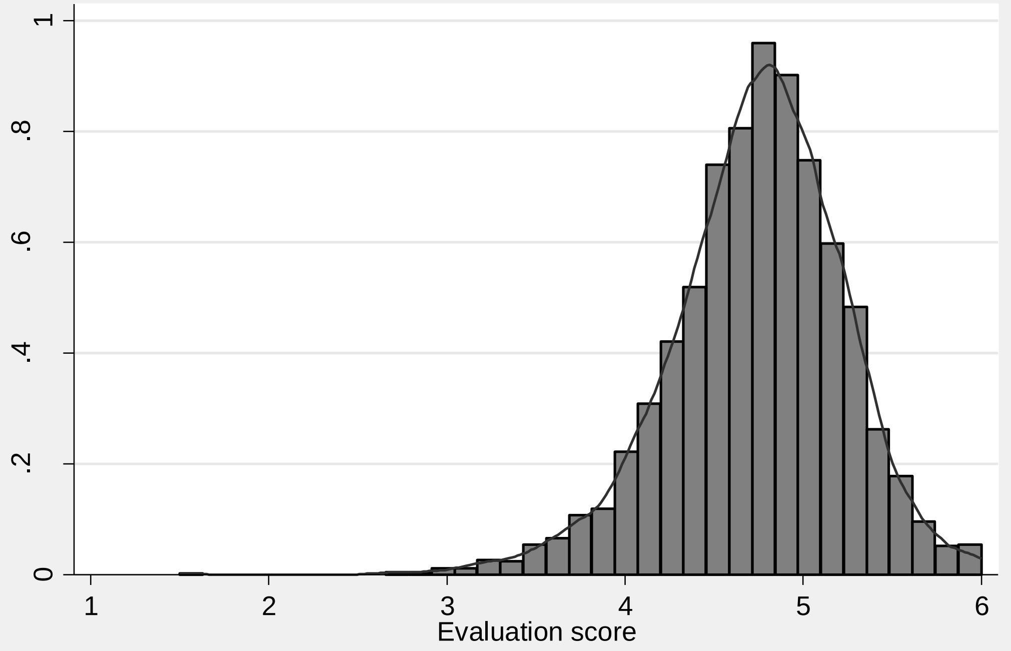
6. Tables and Figures

Figure 1. Evaluation response at the course level



Note: Line is a kernel density estimate.

Figure 2. Evaluation score at the course level



Note: Line is a kernel density estimate.

Figure 3. Selection bias and the response rate

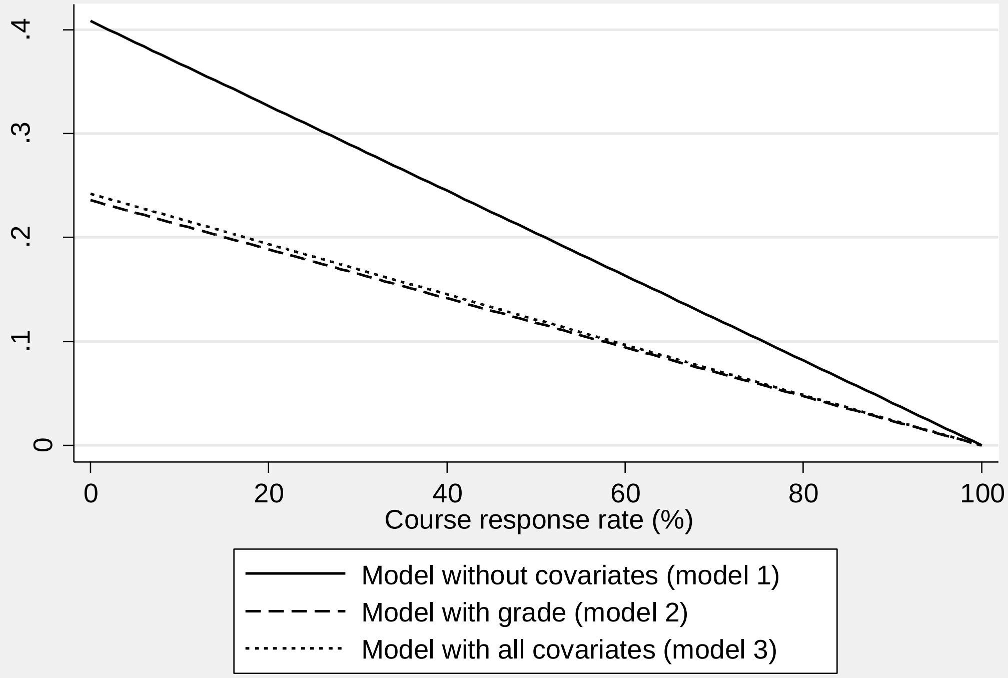
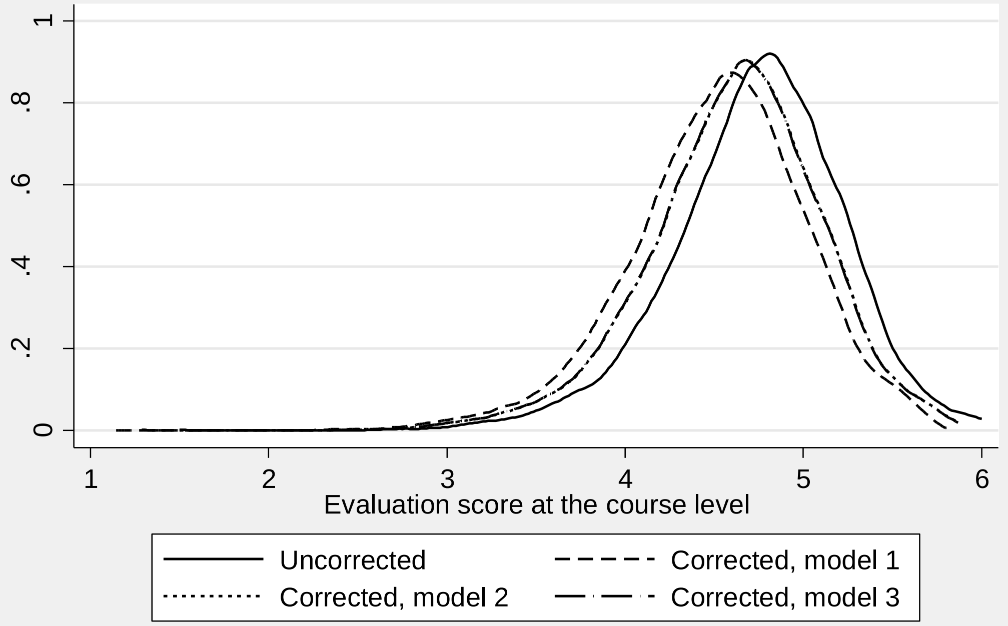


Figure 4. Corrected and uncorrected course evaluations



Note: Lines are kernel density estimates. Model 1 is without covariates;
Model 2 contains the grade; Model 3 contains all covariates.

Table 1. Evaluation questions

The teacher makes clear what knowledge and skills I should acquire to pass this course.
The examination matches the proposed aims of the course (i.e. matches the the knowledge and skills the teacher states I should acquire).
The teaching method (e.g. lectures, assignments, usage of online learning environment) has helped me prepare for the course examination.
The teaching method (i.e. lectures, tutorials, assignments, …, all taken together) stimulated me to participate actively.
The study materials (slides, online learning environment, …) helped me study the course material.
The program of study / learning track of material taught during contact hours was sufficiently clear and coherent.
The teacher made efforts to make the coure interesting.
The teacher communicates clearly about practical matters and course organization.
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, ..).
I am satisfied with the quality of teaching in this course.
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 2A. Number of unique observations

	I. Overall	II. Science Engineering & Technology	III. Bio-Medical Sciences	IV. Humanities & Social Sciences
Students	28,243	5,802	7,246	17,592
Courses	3,329	1,061	591	1,677
Teachers	1,781	564	484	883
Evaluation questions	160	46	26	88
Degree programs	307	131	67	252
Faculties	14	3	3	8
Student-course-teacher	350,535	63,675	89,609	197,251
Student-course-question	3,230,696	650,593	700,314	1,879,789
Student-course-teacher-question	3,473,911	716,254	740,552	2,017,105

Note: The total number students and teachers is smaller than the sum across the three department groups since any one student can be enrolled in courses from various departments and any one teacher can teach courses in multiple departments. Also, the number of degree programs across the three department groups adds to more than the total since some degree programs can be undertaken in various departments.

Table 2B. Summary statistics

	I. Overall		II. Science Engineering & Technology		III. Bio-Medical Sciences		IV. Humanities & Social Sciences	
Course characteristics:	*Mean*	*Std dev*	*Mean*	*Std dev*	*Mean*	*Std dev*	*Mean*	*Std dev*
Response rate	45.0%	19.4%	43.7%	20.1%	42.8%	17.4%	46.5%	19.4%
Evaluation score	4.74	0.48	4.72	0.47	4.72	0.49	4.76	0.49
Grade (out of 20)	13.02	1.92	13.09	1.80	13.44	1.90	12.84	1.97
Pass rate	86.2%	16.5%	86.3%	15.8%	88.9%	15.5%	85.2%	17.1%
Course size	68.84	96.16	45.56	67.60	88.53	97.72	76.64	107.55
Percentage first semester	45.9%	49.8%	48.0%	50.0%	43.1%	49.6%	45.6%	49.8%
Student characteristics								
Response rate	35.7%	39.2%	35.7%	39.1%	39.7%	42.1%	34.2%	39.0%
Internal margin response rate	95.1%	15.5%	94.5%	15.5%	97.0%	11.6%	94.5%	16.9%
Grade	11.82	3.22	12.07	3.21	12.57	3.19	11.54	3.23
Percentage female	55.4%	49.7%	31.3%	46.4%	66.6%	47.2%	59.7%	49.1%
Nr of evaluated courses taken	8.11	3.35	9.03	3.35	8.21	2.91	7.84	3.45
Teacher characteristic:								
Nr of evaluated courses taught	2.67	2.00	2.81	1.90	2.43	1.80	3.00	2.24

Notes: Science, Engineering & Technology includes faculties of Science, Engineering Science and Bioscience Engineering; Bio-Medical Sciences includes the faculties of Medicine, Pharmaceutical Science, and Kinesiology and Rehabilitation Sciences; and Humanities and Social Science includes the faculties of Theology and Religious Studies, Law, Economics and Business, Social Sciences, Arts, Psychology and Educational Sciences and the Institute of Philosophy. The evaluation score lies between 1 (worst score) to 6 (best score) and the grade lies between 1 and 20, where a 10 or higher represents a pass. The external margin response rate reflects whether a student has responded to a course evaluation; the internal margin response rate reflects what percentage of evaluation questions for any given course are answered.

Table 3. Participation equation: exploratory analysis

dependent variable: dummy for participation in the evaluation

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Grade (stdized)	0.0628***	0.0585***	0.0672***	0.0728***	-0.0005	0.0675***	0.0728***	-0.0003	0.0728***	0.0034**	0.0057***
	(0.0025)	(0.0024)	(0.0024)	(0.0025)	(0.0015)	(0.0024)	(0.0025)	(0.0015)	(0.0000)	(0.0015)	(0.0016)
Course passed	0.0337***	0.0440***	0.0555***	0.0557***	0.0160***	0.0562***	0.0557***	0.0158***	0.0556***	0.0224***	0.0214***
	(0.0048)	(0.0047)	(0.0046)	(0.0046)	(0.0031)	(0.0046)	(0.0046)	(0.0031)	(0.0000)	(0.0030)	(0.0030)
Female student	0.0769***	0.0630***	0.0682***	0.0685***	-	0.0673***	0.0685***	-	0.0686***	-	-
	(0.0048)	(0.0051)	(0.0050)	(0.0050)		(0.0017)	(0.0050)		(0.0000)		
Nr of evaluated courses	0.0348***	0.0392***	0.0449***	0.0469***	-	0.0452***	0.0469***	-	0.0469***	-	-
taken by student (stdized)	(0.0021)	(0.0022)	(0.0022)	(0.0026)		(0.0009)	(0.0026)		(0.0000)		
Nr of evaluated courses	0.0123***	0.0108***	-	0.0107***	0.0112***	-	0.0107***	0.0117***	-	-	0.0108***
taught by teacher (stdized)	(0.0012)	(0.0011)		(0.0009)	(0.0007)		(0.0009)	(0.0007)			(0.0009)
Course size (stdized)	-0.0110***	-0.0052**	0.0137***	-	-0.0171***	0.0135***	-	-0.0205***	-	-0.0065***	-
	(0.0021)	(0.0022)	(0.0027)		(0.0017)	(0.0028)		(0.0017)		(0.0021)	
Course taught in first	0.1222***	0.1197***	0.1045***	-	0.1240***	0.1060***	-	0.1228***	-	0.1167***	-
semester	(0.0033)	(0.0033)	(0.0038)		(0.0034)	(0.0038)		(0.0034)		(0.0037)	
Controls	*none*	*faculty dummies*	*teacher dummies*	*course dummies*	*student dummies*	*faculty & teacher dummies*	*faculty & course dummies*	*faculty & student dummies*	*faculty, teacher &course dummies*	*faculty, student & teacher dummies*	*faculty, student &course dummies*
Observations	350,486	350,486	350,486	350,486	350,486	350,486	350,486	350,486	350,486	350,486	350,486

Notes: Dependent variable observed at the student-course-teacher level; participation defined as answering at least one question of the questionnaire. Estimated coefficients reported, robust standard error in parentheses. Standard error clustered by student. Estimated with OLS. *** p<0.01, ** p<0.05, * p<0.1

Table 4. Evaluation equation: exploratory analysis
dependent variable: evaluation score

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Grade (stdized)	0.1427***	0.1372***	0.1289***	0.1232***	0.2113***	0.1273***	0.1232***	0.2094***	0.1232***	0.1958***	0.1885***
	(0.0045)	(0.0044)	(0.0044)	(0.0044)	(0.0044)	(0.0044)	(0.0044)	(0.0044)	(0.0000)	(0.0041)	(0.0000)
Course passed	0.0493***	0.0757***	0.0838***	0.0940***	0.0721***	0.0878***	0.0940***	0.0728***	0.0934***	0.0842***	0.0882***
	(0.0112)	(0.0110)	(0.0105)	(0.0104)	(0.0102)	(0.0105)	(0.0104)	(0.0102)	(0.0000)	(0.0090)	(0.0000)
Female student	0.0166**	0.0477***	0.0476***	0.0513***	-	0.0533***	0.0513***	-	0.0511***	-	-
	(0.0084)	(0.0088)	(0.0087)	(0.0087)		(0.0022)	(0.0087)		(0.0000)		
Nr of evaluated courses	-0.0112***	-0.0100**	-0.0039	-0.0161***	-	-0.0077***	-0.0161***	-	-0.0162***	-	-
taken by student (stdized)	(0.0039)	(0.0040)	(0.0041)	(0.0051)		(0.0012)	(0.0051)		(0.0000)		
Nr of evaluated courses	0.0219***	0.0138***	-	-0.0027	0.0135***	-	-0.0027	0.0120***	-	-	-0.0025***
taught by teacher	(0.0027)	(0.0027)		(0.0039)	(0.0024)		(0.0039)	(0.0024)			(0.0000)
Course size (stdized)	-0.0519***	-0.0191***	-0.0361***	-	-0.0082*	-0.0195***	-	-0.0053	-	-0.0497***	-
	(0.0039)	(0.0040)	(0.0054)		(0.0048)	(0.0055)		(0.0050)		(0.0066)	
Course taught in first	-0.0150***	-0.0035	-0.0099	-	0.0150***	-0.0096	-	0.0147***	-	0.0227***	-
semester	(0.0058)	(0.0057)	(0.0071)		(0.0057)	(0.0071)		(0.0057)		(0.0070)	
Controls	*none*	*faculty dummies*	*teacher dummies*	*course dummies*	*student dummies*	*faculty & teacher dummies*	*faculty & course dummies*	*faculty & student dummies*	*faculty, teacher &course dummies*	*faculty, student & teacher dummies*	*faculty, student &course dummies*
Observations	1,295,940	1,295,940	1,295,940	1,295,940	1,295,940	1,295,940	1,295,940	1,295,940	1,295,940	1,295,940	1,295,940

Notes: Dependent variable observed at the student-course-teacher-question level. Estimated coefficients reported, robust standard error in parentheses. Standard error clustered by student. Estimated with OLS. *** p<0.01, ** p<0.05, * p<0.1

Table 5. Selection models

instrument: semester in which the course was taught

	(1)	(2)	(3)
		Evaluation equation	
Grade (stdized)	-	0.1869*** (0.0047)	0.1633*** (0.0054)
Course passed	-	-	0.0291*** (0.0103)
Course size (stdized)	-	-	-0.0548*** (0.0039)
Female student	-	-	0.0227*** (0.0087)
Nr of evaluated courses taught by teacher (stdized)	-	-	0.0096*** (0.0012)
Nr of evaluated courses taken by student (stdized)	-	-	-0.0102** (0.0045)
		Participation equation	
Grade (stdized)	-	0.2233*** (0.0051)	0.1706*** (0.0069)
Course passed	-	-	0.1099*** (0.0134)
Course size (stdized)	-	-	-0.0330*** (0.0061)
Female student	-	-	0.2178*** (0.0134)
Nr of evaluated courses taught by teacher (stdized)	-	-	0.0133*** (0.0015)
Nr of evaluated courses taken by student (stdized)	-	-	0.1029*** (0.0061)
Course taught in first semester	0.3220*** (0.0088)	0.3423*** (0.0089)	0.3463*** (0.0090)
Observations	3,473,911	3,473,374	3,473,374
Rho	0.1928*** (0.0236)	0.0777*** (0.0231)	0.0780*** (0.0217)
Likelihood Ratio test statistic	66.68*** (0.000)	11.36*** (0.001)	12.92*** (0.000)

Notes: Heckman selection model with ordered probit outcome equation. Standard errors clustered by student. Instrument: semester in which the course was taught. *** p<0.01, ** p<0.05, * p<0.1

Table 6. Estimated selection bias

instrument: semester in which the course was taught

	(1)	(2)	(3)
Total bias	0.2246	0.1297	0.1332
Bias from observables	-	0.0448	0.0492
Bias from unobservables	0.2246	0.0850	0.0840
Covariates	*None*	*Grade*	*All*
Observations	3,473,911	3,473,374	3,473,374

Notes: Based on Heckman selection models with ordered probit outcome equation. Standard errors clustered by student. "All covariates" are the full set of covariates as used in Table 3, where only the instrument is excluded from the outcome equation. Instrument: semester in which the course was taught.

7. Appendix

This appendix shows results from various robustness analyses. Appendix Tables 1 and 2 present exploratory analyses separately by departmental group, showing the correlates of participation in the evaluation (Table 1) and of the evaluation score conditional on participation (Table 2).

Appendix Table 3A shows estimates of selection models with ordered probit outcome equations when the instrument for participation is the number of evaluated courses taken by the student. Appendix Table 3B shows estimates of selection models with ordered probit outcome equations when both the semester the course is taught in and the number of evaluated courses taken by the student are used as instruments. In all specifications, evidence of significant selection bias is found, as seen from the likelihood ratio test statistic. Appendix Table 4 then shows the estimated selection bias in the course evaluation when the number of evaluated courses taken by the student is used an alternative (Panel I, columns 1-3) or additional instrument (Panel II, columns 4-6) to the semester in which the course took place. In our preferred specification with all covariates, the estimated total selection bias is 0.30 when the number of courses the student took is used as an instrument and 0.17 when both instruments are used. In the main results, reported in the paper, we find a slightly smaller total bias of 0.13. Similar to the main results, we find that bias due to unobservable characteristics make up the majority of the total bias.

Appendix Tables 5A, 5B and 5C show estimated selection bias when the evaluated score is treated as a continuous variable rather than an ordered outcome: i.e. in these models we use a linear outcome model. Panel I (columns 1-3) in each table show Limited-Information Maximum Likelihood (LIML, also known as "two-step") estimates, and panel II (columns 4-6) shows Full-Information Maximum Likelihood (FIML) estimates. Table 5A uses the semester in which the course was taught as an instrument; Table 5B uses the number of evaluated courses taken by the student; and Table 5C uses both instruments. Significant selection bias is found in all specifications, as evidenced from the significance of the mills ratio coefficient for LIML models and likelihood ratio test statistic for FIML models. Appendix Tables 6A, 6B and 6C present the found selection biases for the models estimated in Appendix Tables 5A, 5B and 5C, respectively. Results are similar, with found biases typically slightly smaller, compared to when ordered probit outcome equations are used.

Appendix Table 1. Participation equation: exploratory analysis

dependent variable: dummy for participation in the evaluation

					I. Science, Engineering & Technology						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Grade (stdized)	0.062***	0.063***	0.072***	0.082***	0.005	0.072***	0.082***	0.005*	0.082	0.008***	0.012***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.003)	(0.005)	(0.005)	(0.003)	(0.000)	(0.003)	(0.003)
Course passed	0.027***	0.030***	0.031***	0.028***	0.026***	0.031***	0.028***	0.026***	0.028	0.024***	0.023***
	(0.010)	(0.010)	(0.010)	(0.010)	(0.006)	(0.010)	(0.010)	(0.006)	(0.000)	(0.006)	(0.006)
Female student	0.074***	0.075***	0.085***	0.086***	-	0.085***	0.086***	-	0.086	-	-
	(0.012)	(0.012)	(0.012)	(0.012)		(0.012)	(0.012)		(0.000)		
Nr of evaluated courses taken by student (stdized)	0.049***	0.051***	0.053***	0.052***	-	0.053***	0.052***	-	0.052	-	-
	(0.005)	(0.005)	(0.005)	(0.006)		(0.005)	(0.006)		(0.000)		
Nr of evaluated courses taught by teacher (stdized)	-0.006***	-0.007***	-	0.007***	-0.002	-	0.007***	-0.002	-	-	0.007***
	(0.002)	(0.002)		(0.002)	(0.001)		(0.002)	(0.001)			(0.002)
Course size (stdized)	-0.031***	-0.025***	-0.044***	-	-0.022***	-0.045***	-	-0.026***	-	-0.030***	-
	(0.005)	(0.005)	(0.006)		(0.004)	(0.006)		(0.004)		(0.006)	
Course taught in first semester	0.112***	0.112***	0.126***	-	0.112***	0.124***	-	0.114***	-	0.118***	-
	(0.007)	(0.007)	(0.008)		(0.007)	(0.008)		(0.007)		(0.008)	
Observations	63,672	63,672	63,672	63,672	63,672	63,672	63,672	63,672	63,672	63,672	63,672
					II. Bio-Medical Sciences						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Grade (stdized)	0.068***	0.056***	0.070***	0.072***	-0.004	0.070***	0.072***	-0.004	0.072	-0.005	-0.003
	(0.005)	(0.005)	(0.006)	(0.006)	(0.003)	(0.006)	(0.006)	(0.003)	(0.000)	(0.003)	(0.003)
Course passed	0.053***	0.058***	0.060***	0.058***	0.023***	0.060***	0.058***	0.023***	0.058	0.025***	0.026***
	(0.011)	(0.011)	(0.011)	(0.011)	(0.007)	(0.011)	(0.011)	(0.007)	(0.000)	(0.007)	(0.007)
Female student	0.073***	0.070***	0.073***	0.074***	-	0.074***	0.074***	-	0.074	-	-
	(0.011)	(0.011)	(0.011)	(0.011)		(0.011)	(0.011)		(0.000)		
Nr of evaluated courses taken by student (stdized)	0.009*	0.017***	0.025***	0.032***	-	0.027***	0.032***	-	0.032	-	-
	(0.005)	(0.005)	(0.005)	(0.006)		(0.005)	(0.006)		(0.000)		
Nr of evaluated courses taught by teacher (stdized)	-0.002	0.004*	-	0.009***	0.006***	-	0.009***	0.006***	-	-	0.009***
	(0.002)	(0.002)		(0.001)	(0.001)		(0.001)	(0.001)			(0.001)
Course size (stdized)	0.039***	0.032***	0.041***	-	0.003	0.039***	-	0.003	-	0.011**	-
	(0.005)	(0.005)	(0.006)		(0.003)	(0.006)		(0.003)		(0.005)	
Course taught in first semester	0.091***	0.092***	0.065***	-	0.091***	0.070***	-	0.091***	-	0.081***	-
	(0.007)	(0.007)	(0.009)		(0.007)	(0.009)		(0.007)		(0.009)	
Observations	89,609	89,609	89,609	89,609	89,609	89,609	89,609	89,609	89,609	89,609	89,609
					III. Humanities & Social Sciences						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Grade (stdized)	0.048***	0.053***	0.061***	0.069***	-0.001	0.063***	0.069***	-0.001	0.069	0.005***	0.007***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.002)	(0.000)	(0.002)	(0.002)
Course passed	0.039***	0.046***	0.063***	0.064***	0.010**	0.063***	0.064***	0.010**	0.064	0.020***	0.019***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.004)	(0.006)	(0.006)	(0.004)	(0.000)	(0.004)	(0.004)
Female student	0.074***	0.058***	0.061***	0.061***	-	0.059***	0.061***	-	0.061	-	-
	(0.006)	(0.006)	(0.006)	(0.006)		(0.006)	(0.006)		(0.000)		
Nr of evaluated courses taken by student (stdized)	0.048***	0.049***	0.052***	0.049***	-	0.051***	0.049***	-	0.049	-	-
	(0.003)	(0.003)	(0.003)	(0.003)		(0.003)	(0.003)		(0.000)		
Nr of evaluated courses taught by teacher (stdized)	0.026***	0.020***	-	0.012***	0.017***		0.012***	0.018***	-	-	0.012***
	(0.002)	(0.001)		(0.001)	(0.001)		(0.001)	(0.001)			(0.001)
Course size (stdized)	-0.023***	-0.014***	0.019***	-	-0.019***	0.019***	-	-0.023***	-	-0.005*	-
	(0.003)	(0.003)	(0.004)		(0.002)	(0.004)		(0.002)		(0.003)	
Course taught in first semester	0.142***	0.136***	0.114***	-	0.145***	0.116***	-	0.143***	-	0.129***	-
	(0.004)	(0.004)	(0.005)		(0.004)	(0.005)		(0.004)		(0.005)	
Observations	197,205	197,205	197,205	197,205	197,205	197,205	197,205	197,205	197,205	197,205	197,205
Controls	*none*	*faculty dummies*	*teacher dummies*	*course dummies*	*student dummies*	*faculty & teacher dummies*	*faculty & course dummies*	*faculty & student dummies*	*faculty, teacher &course dummies*	*faculty, student & teacher dummies*	*faculty, student &course dummies*

Notes: Dependent variable observed at the student-course-teacher level; participation defined as answering at least one question of the questionnaire. Estimated coefficients reported, robust standard error in parentheses. Standard error clustered by student. Estimated with OLS; for all specifications, more than 99% of all predicted value lie within the unit interval. *** p<0.01, ** p<0.05, * p<0.1

Appendix Table 2. Evaluation equation: exploratory analysis

dependent variable: evaluation score

	I. Science Engineering & Technology										
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Grade (stdized)	0.1646***	0.1644***	0.1616***	0.1562***	0.2351***	0.1617***	0.1562***	0.2352***	0.1553***	0.2338***	0.2321***
	(0.0108)	(0.0108)	(0.0104)	(0.0103)	(0.0110)	(0.0104)	(0.0103)	(0.0110)	(0.0103)	(0.0099)	(0.0099)
Course passed	0.0459*	0.0450*	0.0576**	0.0647***	0.0525**	0.0574**	0.0647***	0.0523**	0.0677***	0.0646***	0.0694***
	(0.0250)	(0.0250)	(0.0235)	(0.0229)	(0.0228)	(0.0235)	(0.0229)	(0.0228)	(0.0229)	(0.0204)	(0.0197)
Female student	0.0090	0.0103	0.0246	0.0258	-	0.0257	0.0258	-	0.0258	-	-
	(0.0189)	(0.0194)	(0.0187)	(0.0186)		(0.0187)	(0.0186)		(0.0186)		
Nr of evaluated courses taken by student (stdized)	-0.0078	-0.0079	-0.0179*	-0.0224*	-	-0.0182*	-0.0224*	-	-0.0240**	-	-
	(0.0100)	(0.0100)	(0.0104)	(0.0120)		(0.0104)	(0.0120)		(0.0119)		
Nr of evaluated courses taught by teacher (stdized)	0.0269***	0.0272***	-	-0.0174*	0.0348***	-	-0.0174*	0.0344***	-	-	-0.0133
	(0.0058)	(0.0058)		(0.0098)	(0.0054)		(0.0098)	(0.0054)			(0.0093)
Course size (stdized)	-0.0449***	-0.0467***	-0.0392***	-	-0.1020***	-0.0384***	-	-0.1024***	-	-0.0198	-
	(0.0093)	(0.0098)	(0.0124)		(0.0139)	(0.0126)		(0.0139)		(0.0180)	
Course taught in first semester	-0.0300**	-0.0301**	-0.0160	-	0.0044	-0.0176	-	0.0052	-	-0.0278*	-
	(0.0132)	(0.0132)	(0.0164)		(0.0138)	(0.0164)		(0.0139)		(0.0163)	
Observations	268,463	268,463	268,463	268,463	268,463	268,463	268,463	268,463	268,463	268,463	268,463
	II. Bio-Medical Sciences										
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Grade (stdized)	0.1347***	0.1322***	0.1190***	0.1204***	0.2439***	0.1188***	0.1204***	0.2440***	0.1209***	0.2038***	0.2057***
	(0.0096)	(0.0097)	(0.0099)	(0.0102)	(0.0094)	(0.0100)	(0.0102)	(0.0094)	(0.0102)	(0.0086)	(0.0087)
Course passed	0.1240***	0.1253***	0.0932***	0.1035***	0.1259***	0.0933***	0.1035***	0.1259***	0.1007***	0.0897***	0.0970***
	(0.0252)	(0.0252)	(0.0240)	(0.0239)	(0.0230)	(0.0240)	(0.0239)	(0.0230)	(0.0238)	(0.0200)	(0.0196)
Female student	0.0718***	0.0716***	0.0701***	0.0636***	-	0.0699***	0.0636***	-	0.0620***	-	-
	(0.0176)	(0.0176)	(0.0175)	(0.0174)		(0.0175)	(0.0174)		(0.0173)		
Nr of evaluated courses taken by student (stdized)	-0.0039	-0.0018	0.0109	-0.0106	-	0.0107	-0.0106	-	-0.0104	-	-
	(0.0080)	(0.0081)	(0.0092)	(0.0137)		(0.0092)	(0.0137)		(0.0135)		
Nr of evaluated courses taught by teacher (stdized)	-0.0081*	-0.0066	-	0.0020	-0.0109**	-	0.0020	-0.0108**	-	-	0.0006
	(0.0048)	(0.0049)		(0.0066)	(0.0045)		(0.0066)	(0.0045)			(0.0065)
Course size (stdized)	-0.0299***	-0.0313***	-0.0170*	-	-0.0451***	-0.0172	-	-0.0451***	-	-0.0300**	-
	(0.0070)	(0.0071)	(0.0103)		(0.0082)	(0.0105)		(0.0082)		(0.0144)	
Course taught in first semester	-0.0577***	-0.0571***	0.0087	-	-0.0387***	0.0085	-	-0.0387***	-	0.0861***	-
	(0.0111)	(0.0111)	(0.0158)		(0.0112)	(0.0159)		(0.0112)		(0.0161)	
Observations	307,228	307,228	307,228	307,228	307,228	307,228	307,228	307,228	307,228	307,228	307,228
	III. Humanities & Social Sciences										
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Grade (stdized)	0.1624***	0.1571***	0.1471***	0.1379***	0.2270***	0.1451***	0.1379***	0.2240***	0.1382	0.2179***	0.2020***
	(0.0070)	(0.0067)	(0.0066)	(0.0067)	(0.0068)	(0.0066)	(0.0067)	(0.0068)	(0.0000)	(0.0063)	(0.0063)
Course passed	0.0300**	0.0706***	0.0867***	0.0990***	0.0718***	0.0916***	0.0990***	0.0731***	0.0979	0.0868***	0.0945***
	(0.0145)	(0.0140)	(0.0134)	(0.0132)	(0.0132)	(0.0133)	(0.0132)	(0.0131)	(0.0000)	(0.0116)	(0.0115)
Female student	-0.0078	0.0514***	0.0496***	0.0560***	-	0.0556***	0.0560***	-	0.0563	-	-
	(0.0118)	(0.0119)	(0.0118)	(0.0118)		(0.0119)	(0.0118)		(0.0000)		
Nr of evaluated courses taken by student (stdized)	-0.0148**	-0.0186***	-0.0075	-0.0179***	-	-0.0129**	-0.0179***	-	-0.0175	-	-
	(0.0058)	(0.0057)	(0.0058)	(0.0069)		(0.0059)	(0.0069)		(0.0000)		
Nr of evaluated courses taught by teacher (stdized)	0.0333***	0.0180***	-	-0.0017	0.0162***	-	-0.0017	0.0139***	-	-	-0.0015
	(0.0035)	(0.0035)		(0.0052)	(0.0031)		(0.0052)	(0.0031)			(0.0050)
Course size (stdized)	-0.0608***	-0.0016	-0.0220***	-	0.0214***	0.0006	-	0.0239***	-	-0.0597***	-
	(0.0055)	(0.0056)	(0.0081)		(0.0065)	(0.0082)		(0.0068)		(0.0095)	
Course taught in first semester	0.0087	0.0299***	-0.0113	-	0.0502***	-0.0063	-	0.0495***	-	0.0354***	-
	(0.0078)	(0.0076)	(0.0093)		(0.0075)	(0.0093)		(0.0075)		(0.0093)	
Observations	720,249	720,249	720,249	720,249	720,249	720,249	720,249	720,249	720,249	720,249	720,249
Controls	*none*	*faculty dummies*	*teacher dummies*	*course dummies*	*student dummies*	*faculty & teacher dummies*	*faculty & course dummies*	*faculty & student dummies*	*faculty, teacher &course dummies*	*faculty, student & teacher dummies*	*faculty, student &course dummies*

Notes: Dependent variable observed at the student-course-teacher-question level. Estimated coefficients reported, robust standard error in parentheses. Standard error clustered by student. Estimated with OLS; for all specifications, more than 99% of all predicted value lie within the interval of 1 to 6. *** p<0.01, ** p<0.05, * p<0.1

Appendix Table 3A. Selection models

instrument: nr of evaluated courses taken by student

	(1)	(2)	(3)
		Evaluation equation	
Grade (stdized)	-	0.2023***	0.1743***
		(0.0078)	(0.0064)
Course passed	-	-	0.0373***
			(0.0109)
Course size (stdized)	-	-	-0.0566***
			(0.0040)
Female student	-	-	0.0386***
			(0.0104)
Nr of evaluated courses taught by teacher (stdized)	-	-	0.0104***
			(0.0012)
Course taught in first semester	-	-	0.0246**
			(0.0121)
		Selection model	
Grade (stdized)	-	0.2140***	0.1707***
		(0.0050)	(0.0069)
Course passed	-	-	0.1097***
			(0.0134)
Course size (stdized)	-	-	-0.0330***
			(0.0061)
Female student	-	-	0.2177***
			(0.0134)
Nr of evaluated courses taught by teacher (stdized)	-	-	0.0133***
			(0.0015)
Nr of evaluated courses taken by student (stdized)	0.1006***	0.0978***	0.3462***
	(0.0060)	(0.0060)	(0.0090)
Course taught in first semester	-	-	0.1030***
			(0.0061)
Observations	3,473,911	3,473,374	3,473,374
Rho	0.1801***	0.1905***	0.1846***
	(0.0521)	(0.0561)	(0.0450)
Likelihood Ratio test statistic	11.95***	11.53***	16.80***
	(0.001)	(0.001)	(0.000)

Notes: Heckman selection model with ordered probit outcome equation. Standard errors clustered by student. Instrument: semester in which the course was taught. *** p<0.01, ** p<0.05, * p<0.1

Appendix Table 3B. Selection models

instruments: semester & nr of evaluated courses taken by student

	(1)	(2)	(3)
		Evaluation equation	
Grade (stdized)	-	0.1894***	0.1670***
		(0.0045)	(0.0053)
Course passed	-	-	0.0296***
			(0.0103)
Course size (stdized)	-	-	-0.0553***
			(0.0040)
Female student	-	-	0.0277***
			(0.0086)
Nr of evaluated courses taught by teacher (stdized)	-	-	0.0097*** (0.0012)
		Selection model	
Grade (stdized)	-	0.2226***	0.1706***
		(0.0051)	(0.0069)
Course passed	-	-	0.1098***
			(0.0134)
Course size (stdized)	-	-	-0.0330***
			(0.0061)
Female student	-	-	0.2178***
			(0.0134)
Nr of evaluated courses taught by teacher (stdized)	-	-	0.0133*** (0.0015)
Nr of evaluated courses taken by student (stdized)	0.1028*** (0.0007)	0.0999*** (0.0060)	0.1032*** (0.0061)
Course taught in first semester	0.3250***	0.3443***	0.3460***
	(0.0014)	(0.0090)	(0.0090)
Observations	3,473,911	3,473,374	3,473,374
Rho	0.1869***	0.0975***	0.1093***
	(0.0069)	(0.0213)	(0.0208)
Likelihood Ratio test statistic	738.20***	20.91***	27.60***
	(0.000)	(0.000)	(0.000)

Notes: Heckman selection model with ordered probit outcome equation. Standard errors clustered by student. Instrument: semester in which the course was taught. *** p<0.01, ** p<0.05, * p<0.1

Appendix Table 4. Estimated selection bias

I. instrument: nr of evaluated courses taken by student			
	(1)	(2)	(3)
Total bias	0.2109	0.2683	0.3042
Bias from observables	-	0.0507	0.0596
Bias from unobservables	0.2109	0.2175	0.2446
Covariates	*None*	*Grade only*	*All*
Observations	3,473,911	3,473,374	3,473,374
II. instruments: semester & nr of evaluated courses taken by student			
	(4)	(5)	(6)
Total bias	0.2246	0.1524	0.1712
Bias from observables	-	0.0457	0.0523
Bias from unobservables	0.2246	0.1067	0.1189
Covariates	*None*	*Grade only*	*All*
Observations	3,473,911	3,473,374	3,473,374

Notes: Heckman selection model with ordered probit outcome equation. Standard errors clustered by student. "All covariates" are the full set of covariates as used in Table 3, where only the instrument(s) is (are) excluded from the outcome equation. *** p<0.01, ** p<0.05, * p<0.1

Appendix Table 5A. Selection models

instrument: semester

	I. LIML estimator			II. FIML estimator		
	(1)	(2)	(3)	(4)	(5)	(6)
	Evaluation equation			*Evaluation equation*		
Grade (stdized)	-	0.2000***	0.1683***	-	0.1970***	0.1657***
		(0.0017)	(0.0020)		(0.0043)	(0.0052)
Course passed	-	-	0.0547***	-	-	0.0536***
			(0.0041)			(0.0113)
Course size (stdized)	-	-	-0.0563***	-	-	-0.0557***
			(0.0012)			(0.0041)
Female student	-	-	0.0262***	-	-	0.0229***
			(0.0025)			(0.0086)
Nr of evaluated courses taught by teacher (stdized)	-	-	0.0106***	-	-	0.0104***
			(0.0005)			(0.0012)
Nr of evaluated courses taken by student (stdized)	-	-	-0.0078***	-	-	-0.0094**
			(0.0013)			(0.0044)
	Participation equation			*Participation equation*		
Grade (stdized)	-	0.2233***	0.1705***	-	0.2233***	0.1705***
		(0.0007)	(0.0012)		(0.0051)	(0.0069)
Course passed	-	-	0.1100***	-	-	0.1100***
			(0.0027)			(0.0134)
Course size (stdized)	-	-	-0.0329***	-	-	-0.0330***
			(0.0008)			(0.0061)
Female student	-	-	0.2178***	-	-	0.2178***
			(0.0014)			(0.0134)
Nr of evaluated courses taught by teacher (stdized)	-	-	0.0133***	-	-	0.0133***
			(0.0003)			(0.0015)
Nr of evaluated courses taken by student (stdized)	0.3220***	0.3423***	0.1029***	0.3226***	0.3424***	0.1029***
	(0.0014)	(0.0014)	(0.0007)	(0.0088)	(0.0089)	(0.0061)
Course taught in first semester	-	-	0.3463***	-	-	0.3464***
			(0.0014)			(0.0090)
Observations	3,473,911	3,473,374	3,473,374	3,473,911	3,473,374	3,473,374
Mills ratio (LIML) or rho (FIML)	0.1976***	0.0635***	0.0645***	0.0367***	0.0359***	0.1007***
	(0.0093)	(0.0087)	(0.0087)	(0.0125)	(0.0120)	(0.0102)
Likelihood Ratio test statistic	-	-	-	106.0***	8.645***	8.933***
				(0.0000)	(0.0033)	(0.0028)

Notes: Heckman selection model with linear outcome equation. Standard errors clustered by student in FIML models. Instrument: semester in which the course was taught. *** p<0.01, ** p<0.05, * p<0.1

Appendix Table 5B. Selection models

instrument: nr of evaluated courses taken by student

	I. LIML estimator			**II. FIML estimator**		
	(1)	(2)	(3)	(4)	(5)	(6)
	Evaluation equation			*Evaluation equation*		
Grade (stdized)	-	0.2110***	0.1812***	-	0.1981***	0.1679***
		(0.0027)	(0.0025)		(0.0046)	(0.0053)
Course passed	-	-	0.0640***	-	-	0.0530***
			(0.0044)			(0.0113)
Course size (stdized)	-	-	-0.0588***	-	-	-0.0559***
			(0.0012)			(0.0042)
Female student	-	-	0.0429***	-	-	0.0256***
			(0.0031)			(0.0086)
Nr of evaluated courses taught by teacher (stdized)	-	-	0.0116***	-	-	0.0103***
			(0.0005)			(0.0012)
Course taught in first semester	-	-	0.0264***	-	-	-0.0016
			(0.0043)			(0.0067)
	Participation equation			*Participation equation*		
Grade (stdized)	-	0.2139***	0.1705***	-	0.2139***	0.1705***
		(0.0007)	(0.0012)		(0.0050)	(0.0069)
Course passed	-	-	0.1100***	-	-	0.1099***
			(0.0027)			(0.0134)
Course size (stdized)	-	-	-0.0329***	-	-	-0.0330***
			(0.0008)			(0.0061)
Female student	-	-	0.2178***	-	-	0.2178***
			(0.0014)			(0.0134)
Nr of evaluated courses taught by teacher (stdized)	-	-	0.0133***	-	-	0.0133***
			(0.0003)			(0.0015)
Nr of evaluated courses taken by student (stdized)	0.1005***	0.0979***	0.1029***	0.1006***	0.0980***	0.1030***
	(0.0007)	(0.0007)	(0.0007)	(0.0060)	(0.0060)	(0.0061)
Course taught in first semester	-	-	0.3463***	-	-	0.3463***
			(0.0014)			(0.0090)
Observations	3,473,911	3,473,374	3,473,374	3,473,911	3,473,374	3,473,374
Mills ratio (LIML) or rho (FIML)	0.1525***	0.1346***	0.1756***	0.0417***	0.0491***	0.0986***
	(0.0161)	(0.0165)	(0.0159)	(0.0146)	(0.0127)	(0.0096)
Likelihood Ratio test statistic	-	-	-	14.19***	8.174***	14.96***
				(0.0002)	(0.0043)	(0.0001)

Notes: Heckman selection model with linear outcome equation. Standard errors clustered by student in FIML models. Instrument: number of evaluated courses taken by the student. *** p<0.01, ** p<0.05, * p<0.1

Appendix Table 5C. Selection models

instruments: semester & nr of evaluated courses taken by student

	I. LIML estimator			II. FIML estimator		
	(1)	(2)	(3)	(4)	(5)	(6)
	Evaluation equation			*Evaluation equation*		
Grade (stdized)	-	0.2018***	0.1715***	-	0.1985***	0.1682***
		(0.0016)	(0.0019)		(0.0043)	(0.0052)
Course passed	-	-	0.0552***	-	-	0.0534***
			(0.0041)			(0.0113)
Course size (stdized)	-	-	-0.0568***	-	-	-0.0560***
			(0.0012)			(0.0041)
Female student	-	-	0.0303***	-	-	0.0261***
			(0.0024)			(0.0086)
Nr of evaluated courses taught by teacher (stdized)	-	-	0.0107***	-	-	0.0103***
			(0.0005)			(0.0012)
	Participation equation			*Participation equation*		
Grade (stdized)	-	0.2225***	0.1705***	-	0.2225***	0.1705***
		(0.0007)	(0.0012)		(0.0051)	(0.0069)
Course passed	-	-	0.1100***	-	-	0.1099***
			(0.0027)			(0.0134)
Course size (stdized)	-	-	-0.0329***	-	-	-0.0330***
			(0.0008)			(0.0061)
Female student	-	-	0.2178***	-	-	0.2178***
			(0.0014)			(0.0134)
Nr of evaluated courses taught by teacher (stdized)	-	-	0.0133***	-	-	0.0133***
			(0.0003)			(0.0015)
Nr of evaluated courses taken by student (stdized)	0.1028***	0.0997***	0.1029***	0.1028***	0.0998***	0.1030***
	(0.0007)	(0.0007)	(0.0007)	(0.0061)	(0.0060)	(0.0061)
Course taught in first semester	0.3249***	0.3445***	0.3463***	0.3255***	0.3446***	0.3463***
	(0.0014)	(0.0014)	(0.0014)	(0.0088)	(0.0090)	(0.0090)
Observations	3,473,911	3,473,374	3,473,374	3,473,911	3,473,374	3,473,374
Mills ratio (LIML) or rho (FIML)	0.1838***	0.0775***	0.0898***	0.1007***	0.0462***	0.0516***
	(0.0079)	(0.0077)	(0.0076)	(0.0102)	(0.0122)	(0.0116)
Likelihood Ratio test statistic	-	-	-	96.94***	14.33***	19.87***
				(0.0000)	(0.0002)	(0.0000)

Notes: Heckman selection model with linear outcome equation. Standard errors clustered by student in FIML models. Instruments: semester in which the course was taught and number of evaluated courses taken by the student. *** p<0.01, ** p<0.05, * p<0.1

Appendix Table 6A. Estimated selection bias

instrument: semester in which the course was taught

I. Limited Information Maximum Likelihood (LIML) estimator			
	(1)	(2)	(3)
Total bias	0.1980	0.1034	0.1083
Bias from observables	-	0.0412	0.0459
Bias from unobservables	0.1980	0.0622	0.0624
Covariates	*None*	*Grade only*	*All*
Observations	3,473,911	3,473,374	3,473,374
II. Full Information Maximum Likelihood (FIML) estimator			
	(4)	(5)	(6)
Total bias for the average course	0.1163	0.0825	0.0853
Bias from observables	-	0.0406	0.0448
Bias from unobservables	0.1163	0.0419	0.0405
Covariates	*None*	*Grade only*	*All*
Observations	3,473,911	3,473,374	3,473,374

Notes: Linear outcome models. Standard errors clustered by student. "All covariates" are the full set of covariates as used in Table 3, where only the instrument(s) is (are) excluded from the outcome equation. Instrument: semester in which the course was taught.

Appendix Table 6B. Estimated selection bias

instrument: nr of evaluated courses taken by student

I. Limited Information Maximum Likelihood (LIML) estimator			
	(1)	(2)	(3)
Total bias	0.1540	0.1765	0.2356
Bias from observables	-	0.0435	0.0518
Bias from unobservables	0.1540	0.1330	0.1838
Covariates	*None*	*Grade only*	*All*
Observations	3,473,911	3,473,374	3,473,374
II. Full Information Maximum Likelihood (FIML) estimator			
	(4)	(5)	(6)
Total bias for the average course	0.0678	0.0889	0.1068
Bias from observables	-	0.0409	0.0467
Bias from unobservables	0.0678	0.0481	0.0602
Covariates	*None*	*Grade only*	*All*
Observations	3,473,374	3,473,374	3,473,374

Notes: Linear outcome models. Standard errors clustered by student. "All covariates" are the full set of covariates as used in Table 3, where only the instrument(s) is (are) excluded from the outcome equation. Instrument: semester in which the course was taught.

Appendix Table 6C. Estimated selection bias

instruments: semester & nr of evaluated courses taken by student

I. Limited Information Maximum Likelihood (LIML) estimator			
	(1)	(2)	(3)
Total bias	0.1833	0.1171	0.1345
Bias from observables	-	0.0416	0.0476
Bias from unobservables	0.1833	0.0755	0.0869
Covariates	*None*	*Grade only*	*All*
Observations	3,473,911	3,473,374	3,473,374

II. Full Information Maximum Likelihood (FIML) estimator			
	(4)	(5)	(6)
Total bias for the average course	0.1183	0.0933	0.1046
Bias from observables	-	0.0409	0.0465
Bias from unobservables	0.1183	0.0524	0.0581
Covariates	*None*	*Grade only*	*All*
Observations	3,473,911	3,473,374	3,473,374

Notes: Linear outcome models. Standard errors clustered by student. "All covariates" are the full set of covariates as used in Table 3, where only the instrument(s) is (are) excluded from the outcome equation. Instrument: semester in which the course was taught.