Bertoni, Marco; Rettore, Enrico; Rocco, Lorenzo

Working Paper

If (My) 6 Was (Your) 9: Reporting Heterogeneity in Student Evaluations of Teaching

IZA Discussion Papers, No. 13565

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Bertoni, Marco; Rettore, Enrico; Rocco, Lorenzo (2020) : If (My) 6 Was (Your) 9: Reporting Heterogeneity in Student Evaluations of Teaching, IZA Discussion Papers, No. 13565, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/224007

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13565

If (My) 6 Was (Your) 9: Reporting Heterogeneity in Student Evaluations of Teaching

Marco Bertoni
Enrico Rettore
Lorenzo Rocco

AUGUST 2020

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13565

If (My) 6 Was (Your) 9: Reporting Heterogeneity in Student Evaluations of Teaching

Marco Bertoni
University of Padova and IZA

Lorenzo Rocco
University of Padova and IZA

Enrico Rettore
University of Padova and IZA

AUGUST 2020

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

If (My) 6 Was (Your) 9: Reporting Heterogeneity in Student Evaluations of Teaching[1]

Student Evaluations of Teaching (SET) are subjective measures of student satisfaction that are often used to assess teaching quality. In this paper, we show that heterogeneity in students' reporting styles challenges SET validity. Using administrative data that allow us to track all evaluations produced by each student, we are able to isolate student-specific reporting scales. We show that reporting heterogeneity explains at least one third of the within-course variation in SET.We also document that students sort across elective courses according to their reporting style. As a result, the average evaluation of two otherwise identical electives can differ only because of heterogeneity in the reporting style of students attending them. Using a simulation exercise, we show that this type of sorting coupled with large sampling variability severely alter the ranking of courses within a major, calling into question the use of SET to incentivise teachers.

JEL Classification: I23, I28, D63

Keywords: student evaluations of teaching, reporting heterogeneity, selection

Corresponding author:
Lorenzo Rocco
Department of Economics and Management
University of Padova
via del Santo 33
35123 Padova
Italy
E-mail: lorenzo.rocco@unipd.it

[1] Acknowledgments and disclaimers are reported at the end of the main text.

Introduction

Student Evaluations of Teaching (SET) were introduced at Harvard and the University of Washington back in 1920's by Edwin Guthrie, a psychologist, with the aim of providing feedback to teachers about their teaching practices. Since then, SET have spread all over the world, and today it is hard to find a college where SET are not collected on a regular basis. Their purpose has also broadened. Nowadays, SET are considered by deans, school managers and other stakeholders as a tool to monitor "customer satisfaction", and are often listed among the elements used to decide promotions and hiring.

Hence, it is not surprising that the validity of SET has been put under scrutiny by scholars. A number of studies have concluded that student evaluations can be manipulated by teachers, are biased by non-response and, most notably, do not reflect exclusively teaching effectiveness, but also other factors such as students' expected grade, gender, and the physical appearance of teachers.

In this paper, we take advantage of panel data of SET from a large Italian university, which allows to *track all the evaluations provided by each student*, and we tackle the issue of SET validity from a different perspective. We develop an intuition of Stark and Freishtat (2014), who argue that students might adopt different subjective scales when they rate their teachers. For instance, two students both judging as "fair" a given course might rate it differently if the first thinks that a grade of 6 out of 10 corresponds with a "fair" evaluation, while, according to the second, a grade of 9 is more appropriate to evaluate the same experience. Also, a course can be rated differently by two students if one systematically rates all courses between 6 and 10, while the other grades between 1 and 10.

Reporting heterogeneity may have pervasive implications for the comparability of evaluations of courses attended by different students. Consider for instance the problem of using SET to rank course quality within a degree program to award a "teacher of the year" prize. If students are randomly distributed across courses, or if all courses must be attended by all students, then the distribution of students' reporting styles is the same in all courses and reporting heterogeneity does not bias the relative evaluation of a course.[2] Instead, if students with different reporting styles, say the lenient or the strict, sort into different courses, then reporting heterogeneity will affect the comparability of the average SET across courses. For instance, if

[2] Even in this ideal situation, sampling variability may still threaten the comparison of average SET by course, especially if course size is small.

in a bachelor in managerial economics the lenient self-select into Economics 101 and the strict choose to attend Management 101, the former course will be evaluated higher than the latter even if teaching quality is the same. In this case, the course average SET cannot be used to portray a valid ranking of course quality.

In surveys, the problem of reporting heterogeneity has been addressed by including anchoring vignettes (King et al., 2004). Vignettes are descriptions of common hypothetical situations that respondents are asked to assess. Under the assumption that differences in vignette assessments are only due to differences in reporting styles and that subjects adopt the same reporting style to evaluate the vignettes and their personal conditions, vignette responses can be used to correct self-reports and make them comparable interpersonally.

Although we do not have proper anchoring vignettes for SET, we follow a similar intuition. Our data include students who major in Economics, Law, Engineering and Medicine. Within majors and cohorts, students are further separated in tracks, and students within the same track are offered the same menu of courses. For each major, cohort and track (a stratum, in the sequel), the courses which are attended and evaluated by the large majority of students play the role of vignettes. We refer to the remaining courses as electives.

Since vignette courses are attended and evaluated by approximately all students of a stratum, the average SET are correct estimates of course quality as they are not influenced by sorting. Thanks to the panel structure of the data, we can decompose the total variation in vignette evaluations in three parts: 1) variation due to systematic differences between-courses; 2) within-course variation due to student-specific reporting styles; and 3) within-course residual variation. We find that at most one third of the total variability in individual SET is attributable to systematic differences between-courses, and spell out some implications of this finding for the reliability of the average SET by course size. The remaining two-thirds of variability in vignette evaluations is within course, and shall be ascribed to student-specific reporting heterogeneity for a proportion ranging between one-fourth and one-half, depending on the major.

We then test whether students' sorting across courses is related to their reporting style. We exploit the observed distribution of students across electives, where sorting is possible, to derive the counterfactual average SET of vignettes that we would have observed if they had been evaluated only by the students who attended a given elective, for each elective. Should sub-groups of students attending different electives provide different average evaluations of

the vignettes, we will take this as evidence of sorting across electives depending on reporting styles. By comparing the factual and the counterfactual evaluations of the same vignette, we test the null hypothesis of no-sorting[3] and reject it in three majors out of four. In the fourth – Medicine – our estimates are too imprecise to reach a firm conclusion.

Finally, we ask to what extent sorting and sampling variability affect the ranking of courses. To answer, we set up a simulation by which we repeatedly draw at random one elective course per stratum, we compute the average SET of each vignette evaluated by the students attending the selected elective, and finally we rank the vignettes accordingly. Except for Engineering, our results show dramatic changes in the ranking of vignettes, depending on the subset of students evaluating them. For instance, in the case of Law – where sorting is more pervasive – a vignette ranked 19 (out of 36) according to the unbiased average SET, can move anywhere between rank 12 and rank 33, depending on the subset of evaluators who is considered. Also, the top ranked vignette can lose as much as 10 positions (out of 36) and turn to be a mid-rank course. Even keeping aside concerns about what features of a course (or of a teacher) students actually evaluate, and what SET actually measure, our results highlight that the combination of reporting heterogeneity and sorting, on top of sampling variability, hampers the comparability of courses' evaluations. The straightforward policy implication is that SET should not be used to incentivise, promote or hire teachers, especially within tournament-like schemes, because they do not accurately reflect the *relative* teaching effectiveness of a scholar.

Rather, universities should try to partly address the problems we have highlighted. On the one hand, the size of courses should be chosen by taking into account also the need of minimising sampling variability, besides other considerations. On the other hand, sufficiently many vignette courses should be offered during the first years to estimate students' reporting styles and correct SET in elective courses.

Unfortunately, in our data we have too few vignettes to properly correct individual evaluations or average SET of elective courses, but we provide a simple procedure to achieve this result and a criterion to decide whether such correction is worth pursuing. By requiring the additional step of estimating elective-specific corrections, this procedure reduces the bias due to reporting heterogeneity at the price of increasing the variability in the estimates of course effects. We

[3] Specifically, in the absence of sorting, the distribution of reporting styles among the attendees of an elective would coincide with the one prevailing in the full population (up to sampling variability). Hence, under independent sorting, the average evaluation of a vignette expressed by each subset of students coincides with the average evaluation of the same vignette in the population (up to sampling error).

show that for roughly two thirds of the electives in our data the mean squared error of the corrected average SET of elective courses exceeds the one associated with the raw average SET. Hence, correcting raw average SET may be more harmful than beneficial.

The rest of the paper is organised as follows. Section 2 summarises the relevant literature. Section 3 describes our data. We present our empirical analysis in Section 4 and the procedure to correct SET in Section 5. Concluding remarks follow.

2. Literature

A vast literature has analysed SET and debated on their reliability and validity. SET aim to measure teaching effectiveness, a concept that is intrinsically difficult to define, and they are used for contrasting purposes, such as providing feedback to help teachers improving their courses, and evaluating teaching in promotion, rewarding or hiring procedures.

In this Section we review a few seminal papers, and refer the reader to the ample survey of Spooren et al. (2013) for a complete account of this debate.

A first line of inquiry debates whether SET capture teaching effectiveness or something else. Whatever the boundaries of the concept of teaching effectiveness, there is little doubt that a teacher is good if his or her students learn well and in depth. The contribution of a teacher to his or her students learning is often referred to as a teacher's value added. Carrell and West (2010) exploit the random assignment of teachers to students at the US Air Force Academy. They find that professors who do better in terms of students' performance in their courses, on average, harm students' performance in more advanced classes. Furthermore, they show that SET are positively correlated with the contemporaneous professor's value added and negatively correlated with the professor's contribution to follow-on test scores. These results confirm previous findings of Weinberg et al. (2009) who first thought of using follow-on courses as indicators of teaching effectiveness, given that scores on follow-on courses cannot be manipulated by the promise of higher grades or by teaching to the test. A related investigation by Braga et al. (2014) on data from Bocconi University – where teachers are randomly assigned to students – also finds that students' evaluation are more positive to professors who contribute less to their performances in the follow-on courses. Boring et al. (2016) use experimental and quasi-experimental data to show that there is no correlation between SET and teaching effectiveness, while SET are correlated to students' grade

expectations and teacher gender. Finally, Hoffmann and Oreopoulos (2009) rely on observational data to conclude that the average SET received by a given instructor over several years and classes predicts student performance more accurately than objective indicators of teaching quality, such as rank, part or full time employment, and salary.

A second strand of literature investigates the determinants of SET, and factors that may bias them. Spooren et al. (2013) review this literature and conclude that SET depend on students, teacher and class characteristics. For instance, there is evidence that teachers' age, gender, race, language background and tenure are correlated with SET. More surprisingly, Hamermesh and Parker (2005) find a correlation between SET and instructors' physical appearance, as rated by a panel of students who looked at instructors' pictures. A similar conclusion is reached by Ponzo and Scoppa (2013) in the Italian context. Other papers investigate whether factors unrelated to teaching quality are reflected into SET. McPherson (2006) finds that SET are influenced by grade expectations, class size, the major chosen by students in class, the semester when the course is offered. However, only in a few cases a professor's rank changes significantly after accounting for these factors. According to Braga et al (2014) even weather conditions prevailing when students evaluate their professors matter. Finally, Hessler et al, 2018, show that the provision of chocolate cookies – a content-unrelated intervention – enhances course evaluations.

Much attention has been recently devoted to whether female teachers receive better or worse evaluation than their male counterparts. Wagner et al. (2016) exploit within-course variation in courses taught by multiple teachers, and find that female teachers are penalised by students. Boring (2017) and Mengel et al. (2018) exploit settings where instructors are randomly assigned to students and confirm this finding.[4]

Another stream of the literature questions the validity of SETs as an accountability tool for teachers. Stark and Freishtat (2014) are very critical against SETs, and especially the common use of comparing the average evaluations of courses within a school. They argue that such averages would be a valid indicator only if SET were genuinely cardinal measures, rather than qualitative judgements arbitrarily associated to numbers, and if all students adopted the same

[4] They also find that such gender bias extends to questions unrelated to teaching, such as how promptly assignments are graded, how good are learning materials and other questions about course organization which are kept constant in the experiment.

scale to express their appreciation for a course.[5] They state that the widespread use of average SET

> *…presumes that the difference between 3 and 4 means the same thing as the difference between 6 and 7 […] that the difference between 3 and 4 means the same thing to different students […] that 5 means the same thing to different students and to students in different courses […] that a 3 "balances" a 7 to make two 5s.*

Using observational data, Goos and Salomons (2017) study non-response bias and suggest that respondents evaluate more generously than non-respondents. Similar results are reported in Treishl and Wolbring (2017) and Spooren and Van Loon (2012).

It is often maintained that students, who do not know the subject taught, can hardly judge teacher's competence (Hornstein, 2017), while SET can be manipulated by an instructor's grading policies (Langbain, 2008), classroom entertainment quotient, and the choice of classroom activities shortly before and on the day of SET administration (Becker and Watts, 1999). Finally, as argued by Braga et al. (2014), students' objectives might be different from those of university administration which uses SET. The former may simply care about their grades, whereas in most cases, the latter care about students' learning.

We contribute to this debate by providing evidence that reporting heterogeneity plays an important role in SET, and by assessing the consequences of reporting heterogeneity for the ranking of courses within a major.

3. Data

3.1. The data and the institutional context

We use administrative data including all SET produced by three cohorts of students matriculated in a large Italian university between October 2011 and October 2013, whom we follow through academic years 2011/12 to 2013/14. We focus on the students enrolled in a bachelor (*laurea triennale*) in Economics; a bachelor in Civil Engineering and its natural continuation, the master (*laurea magistrale*) in Civil Engineering; the five-years degree (*laurea a ciclo unico*) in Law and the six-years degree in Medicine and Surgery. Below, we treat the

[5] Additional negative implications of the ordinal nature of test scores in education are discussed by Bond and Lang, 2013. In our analysis we abstract from such considerations.

bachelor and the master in Civil Engineering as a unique five-years degree since the large majority of bachelors continue to the master.

To hold class size manageable, students in each major are further split in tracks, defined according the initial letter of students' family name, to the location where teaching is actually provided, or to other criteria. The degree in Economics is organized in two tracks, defined on students' family name; the degree in Engineering in a single track, the degree in Law in three tracks, according to students' family name and location of teaching; and the degree in Medicine in four tracks, which reflect students' need to attend practical sessions at the hospital. Fundamental courses, which characterize a major and are compulsory to all students, are offered within each track. Elective courses, that are taken by relatively few students, can either be offered within each track or across tracks, depending on the availability of teachers and instruction rooms.

We combine the organization in tracks of each major and the year of matriculation to partition students in groups, defined by the common feature that all students belonging to a specific group are "at risk" of attending lectures with the same set of teachers in each academic year. We refer to such groups as strata. Accordingly, we define 6 strata for Economics, 3 for Engineering, 9 for Law and 12 for Medicine.

We define a course as a learning unit taught by a specific professor to students belonging to a given stratum.[6] For instance, lectures and tutorials of Economics 101 offered in the academic year 2011/12 to the cohort first matriculated in October 2011, with the initial letter of the last name between A and L, are treated as two separate courses if they are taught by different professors, and as a single course if the same instructor is in charge of both parts. In general, when several teachers are involved in the learning unit, students fill a separate evaluation form for each teacher. We also treat a given learning unit taught by the same professor in two academic years as two separate courses, because the attendees belong to two different cohorts.

In our data we count 201 courses in economics; 79 in engineering; 210 in law and as many as 987 in medicine, where there is a high prevalence of learning units organized in many small sub-units taught by different teachers.

[6] Courses can have very different sizes, ranging from a minimum of 6 hours to a maximum of over 96 hours. For instance, the lectures of Economics 101 amount to 49 hours and the tutorials to 21 hours.

Among these, we define as vignettes the four courses with the highest coverage in each stratum. In most cases, vignettes are compulsory courses offered during the first or second year of the degree. In very few cases vignettes are non-compulsory courses which are very popular and have been chosen by most students. We refer to electives to indicate all courses which are not tagged as vignettes, and the only distinction we make is between vignettes and electives.

The key feature of our dataset is that we can track all the evaluations provided by a specific student. This is possible because students register for exams and fill in SET questionnaires from their electronic personal account on the University's web system, which records their identity. This is needed in order to assure that students evaluate each course only once. Clearly, students' identity is not transmitted to teachers, who only receive aggregate data on their evaluations. To the specific purpose of this project, however, we have been granted access to anonymised micro-level data.

3.2. Sample selection

Starting from the full sample of evaluations provided by students that are present in our data, we apply several selection criteria dictated by the need to have at least three evaluations of vignette courses per student – the minimum number to estimate individual-specific linear reporting scales. As a result, the selected sample turns out to be significantly smaller than the initial one, especially for Law and Medicine. We illustrate the details of our sample selection procedure in the Appendix, where we also provide evidence that the selected sample is not far from being representative of the students attending our four majors. However, the possible lack of representativeness is the unavoidable price to pay to elicit a set of students for whom we observe enough evaluations.

The final sample includes 443 students evaluating 147 courses in Economics,[7] 133 students evaluating 44 courses in Engineering; 477 students and 130 courses in Law; and 339 students and 149 courses in Medicine (Table A1).

Table 1 summarises a few key features of our data. Consistent with the design, each student evaluates at least three and at most four vignettes. The average number of elective courses evaluated is 10.39 in Economics, 13.44 in Engineering, 4.16 in Law and 6.01 in Medicine. A vignette is evaluated, on average, by a number of students ranging between 30.1 in Medicine and 69.54 in Economics, while an elective is evaluated by a number of students ranging

[7] Detailed numbers of elective courses by stratum are reported in Tables A3-A6 in the Appendix.

between 18.7 in Medicine and 49.64 in Engineering. These figures imply that, in the final sample, the coverage rate of vignettes varies between 87 and 96 percent and that of electives between 38 percent in Law and 73 percent in Engineering.

3.3. The SET questionnaire and descriptive evidence

Once they first register online for an exam, students who attended the course and are willing to provide their evaluations are redirected to the SET questionnaire. Attendees are first asked to assess their satisfaction with the following items:

1- Clear presentation of learning objectives from the beginning
2- Clear presentation of the exam rules from the beginning
3- Punctuality of the instructor
4- Quality of lecture notes/reference books
5- Instructor's ability to motivate the class
6- Instructor's ability to teach in a clear way
7- Sufficient prerequisites
8- Workload consistent with ECTs
9- Students' prior interest for the topic of the course

Finally, students are asked to rate their overall satisfaction with the course.[8] The answer to each question is provided along a discrete ordinal scale ranging from 1 to 10. As most literature, we take students' overall satisfaction as the main indicator of SET, akin to the university administration which focuses on overall satisfaction in its official reports.

We analyse the relationship between overall satisfaction and satisfaction with the various aspects of the course rated by students and described above by regressing overall satisfaction on satisfaction with each item, separately for each major. Result reported in Table 2 show that the items more related with overall satisfaction are those capturing teaching effectiveness, followed by personal interest and workload adequacy. The organizational dimensions of the course turn out to play a minor role in determining students' satisfaction. We conclude that "overall satisfaction" is a reasonable one-dimensional index to judge teaching quality.[9] In a

[8] The question students face reads as follows: "Overall, how satisfied are you with this course?"
[9] It is also worth noticing that these items explain a large share of the variation in overall satisfaction, as the R-squared of the regressions is always above 0.7.

robustness analysis, we have depurated "overall satisfaction" from the component depending on "personal interest", to net out taste heterogeneity unrelated to teaching.[10] We anticipate that results do not change qualitatively. In addition, results of these regressions are wholly unaltered when we include students' and teachers' observable characteristics as additional controls.

The average evaluations of the vignettes of each major (pooling all strata) with the corresponding 95% confidence interval are displayed in Figure 1. Considering that close to all students within a track evaluate vignettes, average SET evaluations provide estimates of course quality as measured by overall student satisfaction that are comparable with each other. We notice that average evaluations are rather compressed and - while it is possible to statistically distinguish between top and poor performers - in many cases they do not statistically differ from one another. Considering that vignette courses are the ones evaluated by the largest number of students within a stratum, this result casts concerns about the reliability of the ranks which include small elective courses. We provide more evidence in this sense in the next Section, where we decompose the total variance of vignette courses and spell out implications for inference on average SET by course size.

4. Empirical analysis

Our empirical analysis proceeds in three steps. First, we focus on vignette courses - in which students' sorting can be neglected - and decompose the variability of students' evaluations in three parts, one reflecting systematic differences across courses, one depending on reporting heterogeneity and a one which is residual. This allows us to assess the relevance of reporting heterogeneity and sampling variability. Second, we test the presence of students' sorting on reporting style across elective courses. Finally, we use a simulation exercise to spell out implications of reporting heterogeneity and sampling variability on the ranking of courses which is determined on the basis of their average SET.

In the sequel, for any variable x_{ij}, we denote by $x_{i.}$ the (sample) average of x_{ij} for student i across courses, by $x_{.j}$ the (sample) average of x_{ij} for course j across students, and by $x_{..}$ the overall (sample) average of x_{ij}.

4.1. Variance decomposition for vignette courses

[10] If prior taste was the main driver of satisfaction, we would expect that elective courses received higher evaluations than vignettes. Instead, consistently across all majors, we find that vignettes receive higher evaluations than elective courses, even after including individual-by-semester fixed effects to partly account for selection in elective courses and for learning effects.

Following King et al., 2004, we model student *i* 's evaluation of vignette *j*, denoted y_{ij}, as

$$y_{ij} = \alpha_i + \beta_i \gamma_j + \varepsilon_{ij} \quad (1)$$

where γ_j is a course specific component, ε_{ij} is an individual-by-course component with $\varepsilon_{.j}$ converging to zero as the number of students evaluating course *j* gets larger. Each student reports γ_j upon the student-specific linear transformation (1) plus the noise component, ε_{ij}. The parameter α_i captures how lenient a student is when he or she rates a course,[11] while the parameter β_i captures his/her sensitivity to variations in γ_j. We normalize the averages over students of these parameters, denoted $\alpha_.$ and $\beta_.$, to zero and one respectively, within each major, and we assume that all strata within each major are representative of the student population in the major. We carefully note that the major-specific normalization comes as a straightforward consequence of the impossibility to compare SET across majors due to the lack of any common vignette.

The average evaluation of vignette *j* is $y_{.j} = \alpha_. + \beta_. \gamma_j + \varepsilon_{.j}$. In courses attended by many students, as vignettes are, we can safely approximate $\varepsilon_{.j} = 0$ and, given our normalizations, we get that $y_{.j} = \gamma_j$, that is, the course component γ_j coincides with average student evaluation. This observation helps interpreting model (1): the course component γ_j captures the part of student evaluation which is common across all students and corresponds to their average evaluation of teacher effectiveness, other teacher characteristics, including physical appearance, organizational quality, as well as average student tastes for course *j*. Reporting heterogeneity arises by allowing (α_i, β_i) to vary across students. Each student reporting function is a student-specific linear transformation of the course component γ_j plus a zero-mean residual. The latter includes student *i*'s specific tastes for course *j*, trembling-hand errors in evaluation and random shocks. For simplicity we refer to the ε_{ij} component of student evaluation as noise, to highlight its unsystematic nature.

In a major there are *N* students and *K* vignettes, all strata combined. Students are indexed by *i*=1, ..., N and vignettes by *j*=1, …, K. Each vignette *j* is attended by n_j students and each student *i* evaluates k_i vignettes.

[11] For instance, leniency may be related with individual tastes such as a higher overall interest for or satisfaction with the content of the courses taught in the major.

We start by decomposing the total deviance of y_{ij} in deviance between- and within- course as follows

$$\sum_{j=1}^{K}\sum_{i=1}^{n_j}\left(y_{ij}-y_{..}\right)^2=\sum_{j=1}^{K}n_j\left(y_{.j}-y_{..}\right)^2+\sum_{j=1}^{K}\sum_{i=1}^{n_j}\left(y_{ij}-y_{.j}\right)^2 \quad (2)$$

The first term on the right-hand side of (2) is the between-course deviance and the second is the within-course one. Since for vignette evaluations we can approximate $y_{.j}=\gamma_j$, the between-course variation is unaffected by students' reporting style and reflects only genuine differences between courses. The within-course deviance accounts for the variation among individual SET, and combines reporting heterogeneity and all residual variation (noise). Thanks to the panel nature of our data we can take apart these two factors.

An unbiased estimator of the within-course *variance* is

$$s_u^2=\sum_{j=1}^{K}\frac{n_j-1}{\sum_{j=1}^{K}n_j-1}\left[\frac{\sum_{i=1}^{n_j}\left(y_{ij}-y_{.j}\right)^2}{n_j-1}\right] \quad (3)$$

which accounts for the loss of degrees of freedom involved in the estimate of the course mean $y_{.j}$.

By substitution of (1), expression (3) turns into

$$s_u^2=\sum_{j=1}^{K}\frac{n_j-1}{\sum_{j=1}^{K}n_j-1}\left[\frac{\sum_{i=1}^{n_j}\left(\alpha_i+\beta_i\gamma_j-\gamma_j\right)^2+\sum_{i=1}^{n_j}\varepsilon_{ij}^2}{n_j-1}\right]=$$

$$=\sum_{j=1}^{K}\frac{n_j-1}{\sum_{j=1}^{K}n_j-1}\left[\frac{\sum_{i=1}^{n_j}\left(\alpha_i+\beta_i\gamma_j-\gamma_j\right)^2}{n_j-1}\right]+\sum_{j=1}^{K}\frac{n_j-1}{\sum_{j=1}^{K}n_j-1}\left[\frac{\sum_{i=1}^{n_j}\varepsilon_{ij}^2}{n_j-1}\right] \quad (4)$$

i.e. the sum of variability due to reporting heterogeneity (first term) and due to noise (second term). If reporting heterogeneity was absent, so that $\alpha_i=0$ and $\beta_i=1$ for all students, within-course variability would only depend on noise. We estimate $s_\varepsilon^2=\sum_{j=1}^{K}\frac{n_j-1}{\sum_{j=1}^{K}n_j-1}\left[\frac{\sum_{i=1}^{n_j}\varepsilon_{ij}^2}{n_j-1}\right]$ and we derive the remaining term of (4) by difference.

The estimation of s_ε^2 first requires us to estimate ε_{ij}. To this end, we estimate model (1) on vignette evaluations, separately by major but pooling all strata within a major. Recalling that for each vignette *j* the component γ_j (approximately) corresponds with $y_{.j}$, we regress y_{ij} on a full set of individual dummies and a full set of interactions between γ_j and individual dummies.

The estimated parameters correspond to all α_i and β_i of model (1),[12] that we use to estimate the residuals $\hat{\varepsilon}_{ij} = y_{ij} - \hat{\alpha}_i - \hat{\beta}_i\gamma_j$. Finally, the estimated variance of residuals is

$$s_\varepsilon^2 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{k_i}\hat{\varepsilon}_{ij}^2}{\sum_{i=1}^{N}(k_i-2)} \quad (5)$$

where we take into account that two degrees of freedom per student are lost in the estimation of $\hat{\alpha}_i$ and $\hat{\beta}_i$. Expression (5) is an unbiased estimator of the variance of the noise component in (1).

Table 3 reports the results of this decomposition of the variance. First, we observe that over two thirds of total variance in SET is within-course. Given that the variance of the estimated γ_js is equal to the within-course variance divided by course size, this finding portrays a worrisome picture for the reliability of average SET for small courses. In Appendix Figure A2 we report, for each major, the ratio of the variance of the sample average as a share of the total variance. The Figure shows that this hyperbolic relationship only plateaus at large values of course size. Therefore, sampling variability shall not be neglected when comparing average SET by course, especially for small courses.

Second, reporting heterogeneity accounts for a proportion of the within-course variance that ranges between 25 percent in Medicine and 46 percent in Engineering. Thus, reporting heterogeneity turns out to be a non-negligible source of variation in SET.

4.2. Students' sorting across electives

By definition, students choose what electives they attend. If they sort across courses depending on their reporting style, the average SET of an elective *e* will not coincide with its γ_e, because the average (α_i, β_i) for students evaluating that elective is not equal to (0, 1). In this subsection we illustrate a method to test whether there is sorting on reporting style and to evaluate the size of the resulting bias.

Each elective course *e* is attended by the set of students S_e, and in each stratum we observe a collection of sets S_e that describes how students distribute across available elective courses. Obviously, sets S_e are not disjoint, as students attend several electives, but their union coincides

[12] Estimates of α_i and β_i are very noisy because they are obtained from three or four observations each. To alleviate the consequences of high sample variability, we drop students whose estimated α_i and β_i are in the first and last percentiles, and re-estimate the model on the resulting sample.

with the stratum. Since each vignette j is associated to one specific stratum, we count E_j electives which are offered in the stratum and can be chosen by the students who attend vignette j.

To test for sorting, for all subsets S_e, e=1,..,E_j, we compare the average evaluation of vignette j provided by students in S_e, denoted $y_{.j|e}$, with γ_j. As explained in Section 4.1, the latter is approximately equal to $y_{.j}$.[13] By using model (1), $y_{.j|e} = \alpha_{.|e} + \beta_{.|e}\gamma_j + \varepsilon_{.j|e}$, where $\alpha_{.|e}$, $\beta_{.|e}$, $\varepsilon_{.j|e}$ are the averages of α_i, β_i and ε_{ij} conditional on S_e. If there is sorting, then $\alpha_{.|e}$ and $\beta_{.|e}$ will systematically differ from 0 and 1, respectively. In addition, given the small size of sets S_e, the term $\varepsilon_{.j|e}$ will generally be non-zero. Hence, deviations of $y_{.j|e}$ from γ_j will depend on both sorting and sampling error.

Separately for each major, and pooling strata within majors, Figure 2 reports the scatterplot of $y_{.j|e}$ against the corresponding $y_{.j}$ on the horizontal axis. For each vignette j, there are E_j different $y_{.j|e}$, each one represented by a dot in the figure. The dispersion of $y_{.j|e}$ conditional on $y_{.j}$ is larger for Law, Medicine and Economics and much smaller for Engineering, reflecting differences across majors in the share of students evaluating each elective (the Engineering major offers less electives than the other majors - see Table 1 - and the average size of elective courses is 78 percent of the average vignette size, compared to 54 percent in Economics, 46 percent in Law and 62 percent in Medicine).

In the absence of sorting, the dispersion of $y_{.j|e}$ around γ_j would depend only on noise, and the average of $y_{.j|e}$ across all electives would be equal to γ_j. To test this hypothesis, we regress $y_{.j|e}$ on γ_j and test the null hypothesis of zero intercept and unitary slope. Results are reported in Table 4. In all majors but Medicine, we reject the null of no sorting. In the case of Medicine, estimates are very imprecise and no firm conclusion can be established.

We remark that that there are special instances where sorting on reporting style is compatible with zero intercept and unitary slope. Therefore, our proposed test under-rejects the null. Consider, for instance, a situation in which there are only two electives, *e=1,2*, and students *perfectly and symmetrically* sort half and half between them. Since S_1 is the complementary set of S_2 with respect to the stratum population, and in the population $\alpha_{.} = 0$ and $\beta_{.} = 1$, we

[13] In principle, one could also consider the option of correcting individual students' evaluations of elective course by means of the estimated $\hat{\alpha}_i$ and $\hat{\beta}_i$, derived in the previous section. This is an option we did not take into consideration given the large sampling variability associated with those estimates. We will nonetheless consider a related procedure in Section 5.

have that $y_{.j|1}$ and $y_{.j|2}$ are symmetric with respect to γ_j and the average of $y_{.j|e}$ would coincide with γ_j.[14]

Rejecting the null of no sorting does not necessarily imply that sorting on reporting styles is a concern of practical interest, as a large part of the deviations $y_{.j|e} - y_{.j}$ reported in Figure 2 could still be attributable to sampling variability. To assess how relevant sorting is with respect to noise we proceed as follows.

The deviation $y_{.j|e} - y_{.j}$ can be written as $y_{.j|e} - y_{.j} = \alpha_{.|e} + (\beta_{.|e} - 1)\gamma_j + \varepsilon_{.j|e}$. For each vignette *j*, the average of the squared deviation[15] is

$$\begin{aligned} &\tfrac{1}{\mathrm{E}_j}\textstyle\sum_e (y_{.j|e} - y_{.j})^2 = \tfrac{1}{\mathrm{E}_j}\textstyle\sum_e \varepsilon_{.j|e}^2 + \tfrac{1}{\mathrm{E}_j}\textstyle\sum_e (\alpha_{.|e} + (\beta_{.|e} - 1)\gamma_j)^2 = \\ &= \tfrac{1}{\mathrm{E}_j}\textstyle\sum_e \varepsilon_{.j|e}^2 + Var(\alpha_{.|e} + (\beta_{.|e} - 1)\gamma_j) + [E(\alpha_{.|e} + (\beta_{.|e} - 1)\gamma_j)]^2 \end{aligned} \qquad (6)$$

If the allocation of students across electives was random, then $y_{.j|e} \simeq y_{.j} \simeq \gamma_j$, and the term $[E(\alpha_{.|e} + (\beta_{.|e} - 1)\gamma_j)]^2$ would vanish.[16] The latter component is the square of the systematic deviation between $y_{.j|e}$ and γ_j, which only emerges under sorting, and can be estimated by $\left(\mathrm{y}_{.j} - \frac{1}{\mathrm{E}_j}\sum_e y_{.j|e}\right)^2$. Hence, the ratio $S = \frac{\left(\mathrm{y}_{.j} - \frac{1}{\mathrm{E}_j}\sum_e y_{.j|e}\right)^2}{\frac{1}{\mathrm{E}_j}\sum_e (y_{.j|e} - y_{.j})^2}$ is an index, defined between 0 and 1, which measures sorting intensity.[17] Figure A3 in Appendix shows that the average $y_{.j|e}$ does not coincide with $y_{.j}$, and in several cases the deviation is substantial. Next, Figure A4 reports the value of *S*, by major and vignette. On average, *S* is 0.152 in Economics, 0.477 in Engineering, 0.186 in Law and 0.229 in Medicine. In all majors there are vignette for which sorting is predominant and *S* can even exceed 0.80.

4.3. Implications of reporting heterogeneity and noise

[14] In our data, the union of S_e does not coincides with the set of students who evaluate any vignette *j* because in the sample there are students who evaluates only vignettes but not electives (as apparent in the case of stratum 6 in Medicine). Moreover, the number of evaluations expressed by students is not constant and the distribution of evaluations across courses is uneven.

[15] Recall that, with sorting, the average of $y_{.j|e}$ across electives does not coincide with $y_{.j}$ and so the average squared deviation is not the variance of $y_{.j|e}$.

[16] In all cases, the average of $\varepsilon_{.j|e}$ across electives is approximately zero.

[17] Ratio S is a lower-bound for the proportion of the dispersion of $y_{.j|e}$ around $y_{.j}$ which is due to sorting. With sorting, also the component $Var(\alpha_{.|e} + (\beta_{.|e} - 1)\gamma_j)$ of the second line of equation (6) increases.

We now turn to illustrate the combined effect of sorting and noise on the ranking of courses based on average SET. The average SET is the indicator typically used by universities to assign teaching awards (and the connected benefits), or sanction teachers.[18] Therefore, it is of policy relevance to gauge the extent to which such ranking is sensible to the lack of validity and of reliability of SET that we have documented so far.

We focus on vignette courses, the courses for which we can (approximately) observe the true ranking based on γ_j. We compare this ranking to the counterfactual rankings that would result if only the students attending an elective evaluated the vignettes. More precisely, for each stratum t=1,2,..,T_M of a major M, we randomly draw one elective e and we take the corresponding $y_{.j|e}$ for all vignettes evaluated by students in S_e. Next, we sort all these evaluations and derive the corresponding ranking of the vignettes associated to that major. This ranking differs from the true one due to both reporting heterogeneity and the noise component – a non-negligible source of variability given the small size of the elective courses. We repeat this procedure 200 times to derive an empirical distribution of the counterfactual rankings. In Figure 3 we display the boxplot of the ranks that each vignette can take across the 200 replications.

In all majors we observe that the rank of the vignettes with the highest and the lowest average evaluation does not change much across replications. Instead, the rank of mid-range vignettes varies widely. Partly, this result depends on the fact that evaluations are rather compressed, as shown in Figure 1, and even small perturbations produce large variations in rank. For mid-range vignettes, the interquartile range of their ranks can exceed 10 positions in Law and Medicine and 3 or 4 positions in Economics. For Engineering, however, vignette ranks are quite stable. This is not surprising given the small dispersion of $y_{.j|e}$.

We have replicated all our analysis using two alternative measures of course quality. First, to make use of all the information on SET provided by students we have used a principal component analysis to extract a single factor out of all SET questions present in the questionnaire. Second, we have also filtered out "interest for the subject" from overall satisfaction by using the coefficient estimated in Table 3. This avoids to blame teachers for students' low interest in the subjects they teach, and also alleviates concerns related with

[18] The worst performers might be penalized in promotion, sometimes on salary progressions, and when SET are made public by the stigma of colleagues and students.

sorting on taste for the subject as the main driver of our findings. In both cases, we have obtained comparable results (not reported but available from the authors).

5. Correcting SET of elective courses

In theory, the availability of vignette courses offers the opportunity of anchoring students' evaluation and making them comparable. In practice, correcting evaluations requires a preliminary estimation step which increases sampling variance. As a result, the correction could increase rather than reduce our uncertainty about the true SET. In this section we discuss a simple procedure to correct the evaluations of elective courses and a criterion to decide whether correcting is worthwhile.

Applying model (1) to elective course e, $y_{\cdot e} = \alpha_{\cdot|e} + \beta_{\cdot|e}\gamma_e + \varepsilon_{\cdot e}$ is its average evaluation, which combines the ratings of the n_e students who attended it (the set S_e). Parameters $\alpha_{\cdot|e}$ and $\beta_{\cdot|e}$ are the averages of α_i and β_i over the same set of students and $\varepsilon_{\cdot e}$ is the average noise for elective e. Due to sorting, $y_{\cdot e}$ is a biased estimator of γ_e and the bias is

$$E(y_{\cdot e} - \gamma_e) = \alpha_{\cdot|e} + (\beta_{\cdot|e} - 1)\gamma_e \tag{7}$$

We can estimate $\alpha_{\cdot|e}$ and $\beta_{\cdot|e}$ by averaging the evaluations that students in S_e assign to vignette j, which yields $y_{\cdot j|e} = \alpha_{\cdot|e} + \beta_{\cdot|e}\gamma_j + \varepsilon_{\cdot j|e}$, for each j=1,..,4, and regressing $y_{\cdot j|e}$ on γ_j, separately for each e. The latter, as in the previous sections, is approximated by $y_{\cdot j}$, the average evaluation of vignette j taken over all students of a major. Denote $\hat{\alpha}_{\cdot|e}$ and $\hat{\beta}_{\cdot|e}$ the corresponding estimates. Equipped with such estimates, we can compute $\tilde{y}_{\cdot e} = \frac{y_{\cdot e} - \hat{\alpha}_{\cdot|e}}{\hat{\beta}_{\cdot|e}}$. A first order Taylor expansion of $\tilde{y}_{\cdot e}$ around $\alpha_{\cdot|e}$ and $\beta_{\cdot|e}$, for any e, yields

$$\tilde{y}_{\cdot e} \sim \frac{y_{\cdot e} - \alpha_{\cdot|e}}{\beta_{\cdot|e}} - \frac{1}{\beta_{\cdot|e}}\left(\hat{\alpha}_{\cdot|e} - \alpha_{\cdot|e}\right) - \frac{y_{\cdot e} - \alpha_{\cdot|e}}{\beta_{\cdot|e}^2}\left(\hat{\beta}_{\cdot|e} - \beta_{\cdot|e}\right) \tag{8}$$

Up to a first order approximation, $E(\tilde{y}_{\cdot e}) = \gamma_e$, and hence $\tilde{y}_{\cdot e}$ is an approximately unbiased and feasible correction of $y_{\cdot e}$.

As usual, $\hat{\alpha}_{\cdot|e}$ and $\hat{\beta}_{\cdot|e}$ embody sampling variability, which is larger the smaller n_e. Hence, correcting $y_{\cdot e}$ eliminates its systematic bias, but inflates its sampling variability. To assess the trade-off between bias and variability, we compare the mean squared errors of $y_{\cdot e}$ and of its correction $\tilde{y}_{\cdot e}$.

The mean squared error (MSE) associated to the estimator $y_{\cdot e}$ is

$$MSE(y_{\cdot e}) = E(y_{\cdot e} - \gamma_e)^2 = [\alpha_{\cdot e} + (\beta_{\cdot e} - 1)\gamma_e]^2 + \frac{\sigma_\varepsilon^2}{n_e} \tag{9}$$

which is the sum of the squared bias and the sampling variance

$$Var(y_{\cdot e}) = \frac{\sigma_\varepsilon^2}{n_e}. \tag{10}$$

Turning to $\tilde{y}_{\cdot e}$, the mean squared error $E(\tilde{y}_{\cdot e} - \gamma_e)^2$ approximately coincides with the variance of $\tilde{y}_{\cdot e}$, since $\tilde{y}_{\cdot e}$ is approximately unbiased. By using (8), the latter is approximately equal to

$$MSE(\tilde{y}_{\cdot e}) = Var(\tilde{y}_{\cdot e}) = Var\left(\frac{y_{\cdot e} - \alpha_{\cdot|e}}{\beta_{\cdot|e}}\right) + Var\left(\frac{1}{\beta_{\cdot|e}}(\hat{\alpha}_{\cdot|e} - \alpha_{\cdot|e})\right) +$$

$$+Var\left(\frac{y_{\cdot e} - \alpha_{\cdot|e}}{\beta_{\cdot|e}^2}(\hat{\beta}_{\cdot|e} - \beta_{\cdot|e})\right) + 2cov\left(\frac{1}{\beta_{\cdot|e}}(\hat{\alpha}_{\cdot|e} - \alpha_{\cdot|e}), \frac{y_{\cdot e} - \alpha_{\cdot|e}}{\beta_{\cdot|e}^2}(\hat{\beta}_{\cdot|e} - \beta_{\cdot|e})\right) \tag{11}$$

This holds because $cov\left(\frac{y_{\cdot e} - \alpha_{\cdot|e}}{\beta_{\cdot|e}}, \frac{1}{\beta_{\cdot|e}}(\hat{\alpha}_{\cdot|e} - \alpha_{\cdot|e})\right) = 0$ and $cov\left(\frac{y_{\cdot e} - \alpha_{\cdot|e}}{\beta_{\cdot|e}}, \frac{y_{\cdot e} - \alpha_{\cdot|e}}{\beta_{\cdot|e}^2}(\hat{\beta}_{\cdot|e} - \beta_{\cdot|e})\right) = 0$. Intuitively, for given *e*, the sampling error embodied in the estimates $\hat{\alpha}_{\cdot|e}$ and $\hat{\beta}_{\cdot|e}$ is independent of the sampling error embodied in $y_{\cdot e}$, because the former derives from students' evaluation of the vignettes and the latter from students' evaluations of elective *e*.

Using standard formulas for the variance and covariance among the coefficients of the linear regression model, we obtain that

$$MSE(\tilde{y}_{\cdot e}) = \frac{1}{\beta_{\cdot e}^2}\frac{\sigma_\varepsilon^2}{n_e}\left\{\frac{5}{4} + \frac{1}{\sum_{j=1}^4 (y_{\cdot j} - y_{\cdot\cdot})^2}\left((y_{\cdot\cdot} - \gamma_e)^2 + \frac{1}{\beta_{\cdot e}^2}\frac{\sigma_\varepsilon^2}{n_e}\right)\right\} \tag{12}$$

where $y_{\cdot\cdot}$ is the average of $y_{\cdot j}$ over the four vignettes. Comparing (9) and (12) we note that the sampling variance of $\tilde{y}_{\cdot e}$ is certainly larger than that of $y_{\cdot e}$.

It turns out that $MSE(\tilde{y}_{\cdot e})$ is smaller than $MSE(y_{\cdot e})$ only in about 35 percent of elective courses in the major in Economics, 44 percent in Engineering, 37 percent in Law and 35 percent in Medicine. These typically are the electives attended by relatively many students. In the remaining majority of cases, the additional sampling variance brought in by the correction exceeds the systematic bias of $y_{\cdot e}$.

Although in our data there is little gain in filtering out the bias due to sorting, the procedure suggested in this section could be applied more fruitfully in other contexts (e.g. where electives are larger).

Conclusions

Several recent papers have studied whether SET reflect teaching quality or, rather, features that should not affect a fair evaluation of teaching, such as teacher's gender or physical appearance. These studies have exploited experimental settings, where teachers are randomized to students and where sorting is absent.

In this paper we take a different perspective, we are agnostic about the question of what exactly SET measure, and we investigate whether SET are affected by reporting heterogeneity. First, we quantify the proportion of the total variation in SET which results from reporting heterogeneity and from noise. Second, we test whether students sort across courses depending on their reporting style. Third, we document how much the combination of reporting heterogeneity and noise affect the ranking of courses by average SET. Finally, we suggest a procedure to correct SET and a criterion to decide whether to undertake such correction.

The key feature of our dataset is that we can track all evaluations provided by each student. Then, following the logic of the literature using anchoring vignettes (see King et al., 2004), we use courses attended by the large majority of students as vignettes to identify students' reporting styles, and assess how much the average evaluation of a given vignette varies across the sub-groups of students attending each different elective offered in a major.

We find that reporting heterogeneity accounts for one fourth to one third of the within-course variability of SET, which by itself represents about two thirds of total variability of SET. Moreover, we find evidence that sorting on reporting style does exist and it is of practical importance. Sorting on reporting style, jointly with the large variability in the SET due to noise, heavily affect the ranking of courses based on SET: we document many cases of courses that swing between top and bottom ranks as a consequence of the low degree of reliability and validity of SET.

We derive two implications from these results. The first one is that– whatever dimension of teaching quality they measure – SET are neither reliable nor valid. Hence, SET should not be used to decide on a teacher's career. Stark and Freishtat (2014) argue that, at the very minimum,

SET should be accompanied by the evaluation of one or more experts who attend the lectures and are in charge of judging the whole faculty. The second implication is that SET should not be used in comparative evaluations, because the ranking of courses by average SET depends on the peculiar manner students distribute across classes.

While we deem comparisons across majors as absolutely far-stretched, SET could still be useful to compare courses *within a major*, having in mind a couple of crucial caveat: first, they should be made comparable across students; second, they should not be used with courses attended by a small number of students.

How to practically achieve comparability is beyond the scope of this paper. Broadly speaking, comparability could be achieved by introducing in students' curricula a purposively designed set of "vignette courses". These could be courses of general content to be attended and evaluated by all students at the beginning of their career. The evaluations of these vignettes could be used to harmonize SET in all other courses and remove the bias due to reporting heterogeneity and sorting.

Acknowledgements

We thank Martina Miotto and Riccardo Franceschin for excellent research assistance in the early stages of the project. We have benefitted from comments from Giorgio Brunello, Ingo Isphording, Paolo Pinotti, Roberto Nisticò, Ulf Zoelitz, as well as from seminar participants at ANVUR in Rome, Bergamo, ZEW in Mannheim, IRVAPP in Trento, IWAEE 2019 in Catanzaro, Padova, and Verona. We acknowledge financial support from a CARIPARO foundation "Starting Grant". This paper uses confidential data from the archives of a large Italian University. Please contact us for queries on data access and replication. We have no relevant financial or material interest to disclose about this paper. All remaining errors are our own.

References.

Becker, W. E., & Watts, M. (1999). How departments of economics evaluate teaching. *American Economic Review*, 89(2), 344–349.

Bond, T. N., & Lang, K., 2013. The Evolution of the Black-White Test Score Gap in Grades K–3: The Fragility of Results, *Review of Economics and Statistics*, 95(5), 1468-1479

Boring, A., Ottoboni, K., Stark, P.B., 2016. Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research* (DOI: 10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1)

Boring, A. (2017). Gender Biases in Student Evaluations of Teachers. *Journal of Public Economics*, 145, 27–41.

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

Carrell, S.E., West, J.E., 2010. Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy* 118 (3), 409–432.

Hamermesh, D. S. and A. Parker (2005). Beauty in the Classroom: Instructors' Pulchritude and Putative Pedagogical Productivity. *Economics of Education Review*, 24, 369–376.

Hessler, M., Pöpping, D. M., Hollstein, H., Ohlenburg, H., Arnemann, P. H., Massoth, C., Seidel, L. M., Zarbock, A., Wenk, M., 2018. Availability of cookies during an academic course session affects evaluation of teaching. *Medical education*, 52(10), 1064-1072.

Hoffmann, F., Oreopoulos, P., 2009. A professor like me: the influence of instructor gender on college achievement. *Journal of Human Resources* 44 (2), 479–494.

King, G., Murray, C.J.L., Salomon, J.A., Tandon, A., 2004. Enhancing the validity and cross-cultural comparability of measurement in survey research. *APSR*, 98,191–207.

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

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

Mengel, F., Sauermann, J., & Zölitz, U. (2019). Gender bias in teaching evaluations. *Journal of the European Economic Association*, *17*(2), 535-566.

Ponzo, M. & Scoppa, V. (2013). Professors' Beauty, Ability, and Teaching Evaluations in Italy. *The B.E. Journal of Economic Analysis & Policy*, 13(2), pp. 811-835.

Spooren, P. & Van Loon, F. (2012). Who participates (Not)? A non-response analysis on students' evaluations of teaching. *Procedia—social and behavioral sciences*, vol. 69 (pp. 990–996). International Conference on Education and Educational Psychology (ICEEPSY 2012)

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

Stark, P.B., Freishtat, R., 2014. An evaluation of course evaluations. *ScienceOpen Research* (DOI: 10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1)

Wagner, N., Rieger, M., Voorvelt, K., 2016. Gender, ethnicity and teaching evaluations: evidence from mixed teaching teams. *Economics of Education Review* 54 (54), 7994

Weinberg, B. A., Hashimoto, M., & Fleisher, B. M. (2009). Evaluating teaching in higher education. *Journal of Economic Education*, 40(3), 227–261.

Wolbring, T. & Treischl, E. (2015). Selection bias in students' evaluation of teaching. *Research in Higher Education*, 1–21.

Wolbring, T. (2012). Class attendance and students' evaluations of teaching: Do no-shows bias course ratings and rankings? *Evaluation Review*, 36(1), 72–96.

Appendix. Sample Selection.

We start by selecting students who have provided at least one evaluation of teaching as attendees (non-attendees can also evaluate courses, but using a different questionnaire). Evaluations can be missing for two reasons. First, students are asked to evaluate a course when they first register for the final exam, but only if they do so within the academic year in which they attended the course. Late-comers are not permitted to evaluate. Second, and more important, students can refuse to evaluate the course. Nonresponse is common in SET, and in our case is responsible for a large extent of the gap between the number of enrolled students and the size of our reference population. Although excluding non-respondents might introduce a bias, this is not a major concern in this paper, whose purpose is that of documenting the importance of reporting heterogeneity among evaluators.[19]

As reported in Table A1, we retain 598 students in Economics; 242 in Engineering; 1,317 in Law and 953 in Medicine. This is our reference population.

We further refine the sample by dropping students with less than three evaluations, as this is the minimum number of evaluations that we need to estimate student-specific reporting functions. As shown in Table A1, second row, this operation significantly reduces the available number of students for Law, and to a lesser extent for Medicine, Economics and Engineering.

We study reporting heterogeneity in SET by exploiting as anchors those courses that are evaluated by close to all students. We refer to these courses as vignettes. We select four vignettes in each stratum, which correspond to the four courses with the highest coverage. In the bottom panel of Table 1 we report that the average coverage among vignettes respectively reaches 86 and 91 percent in Economics and Engineering, while it is lower for Law and Medicine (67 and 66 percent respectively).

Since we can only rely on vignette responses – not affected by sorting – to estimate student response styles (see below), we further retain only students who evaluate at least three out of the four vignettes defined for their stratum.[20] This requirement implies a substantial reduction in the sample of students, which is necessarily more severe in the degrees of Law and Medicine.

[19] In a few cases students evaluate courses that are supposed to be offered in other strata. This happens more frequently in the majors of Medicine and Law, where students might ask to change track if the timetable or the location of instruction activities fits better with their needs. We drop these students from the sample.

[20] At this stage we apply two additional minor restrictions. First, we drop students whose evaluations of the vignettes are all equal. For them it would not be possible to distinguish the effect of course quality from that of reporting heterogeneity on their evaluation (see below). Second, we drop one stratum in the major of Medicine where the average evaluation is equal among the four vignette.

The number of retained students decreases to 443 in Economics (a 26 percent decline with respect to the reference population), 195 in Engineering (20 percent decline), 477 (64 percent decline) in Law and 405 (58 percent decline) in Medicine.

The large decline in sample size for Law and Medicine raises concerns about the extent to which the retained students are representative of the reference population. To assess possible differences, we test, stratum by stratum, the null of equal average evaluation of vignettes between the students who have evaluated at least 3 vignettes and those who have evaluated less than three vignettes. We reject the null only in two cases out of 30 (i.e. 6.7 percent) - one stratum in Engineering and one in Medicine, which we drop from the sample. This reassuring result is qualitatively confirmed by Figure A1, where we plot the average evaluation of each vignette provided by the sample of students who evaluate at least 3 vignettes and those who evaluate less than 3 vignettes.

We further investigate differences in composition between the reference population and the retained sample in terms of four observable characteristics: gender, the region of birth, the year of birth and the final grade at high school. Results are reported in Table A2, and show that gender is slightly unbalanced in economics and engineering. Overall, however, this analysis suggests that the students in the study sample and in the reference population are comparable to a large extent.

The final step of sample definition regards the elective courses, that is, those courses which do not qualify as vignettes in each stratum. In order to reliably estimate average SET by course, we keep only electives which receive at least ten evaluations.

Eventually, we end up with 443 students evaluating 147 courses in Economics, 133 students evaluating 44 courses in Engineering; 477 students and 130 courses in Law; and 339 students and 149 courses in Medicine. A detailed account of elective courses by stratum is provided in Tables A3-A6.

Tables

Table 1. The study sample – Descriptive statistics. By major.

	Economics	Engineering	Law	Medicine
	(1)	(2)	(3)	(4)
Number of students	443	133	477	339
Number of strata	6	2	9	10
Number of courses				
Vignettes	24	8	36	40
Electives	123	36	94	109
Average number of courses evaluated by each student				
Vignettes	3.77	3.84	3.48	3.55
Electives	10.39	13.44	4.16	6.01
Average number of students evaluating each course				
Vignettes	69.54	63.88	46.17	30.1
Electives	37.44	49.64	21.09	18.7
Coverage (% evaluating)				
Vignettes - at definition	0.86	0.91	0.67	0.66
Vignettes - in final sample	0.94	0.96	0.87	0.89
Electives – in final sample	0.51	0.73	0.38	0.47

Table 2. Overall satisfaction and its covariates. OLS estimates.

	Economics	Engineering	Law	Medicine
Clear presentation of learning objectives from the beginning	0.077***	0.051	0.136***	0.081***
	(0.018)	(0.034)	(0.020)	(0.018)
Clear presentation of the exam rules from the beginning	0.049***	0.069**	-0.028	0.016
	(0.015)	(0.028)	(0.018)	(0.016)
Punctuality of the instructor	-0.003	0.053**	0.027	0.042***
	(0.015)	(0.026)	(0.017)	(0.016)
Quality of lecture notes/reference books	0.078***	0.099***	0.056***	0.069***
	(0.013)	(0.021)	(0.017)	(0.016)
Instructor is able to motivate the class	0.213***	0.195***	0.228***	0.319***
	(0.017)	(0.034)	(0.019)	(0.019)
Instructor teaches in a clear way	0.284***	0.342***	0.275***	0.252***
	(0.017)	(0.032)	(0.022)	(0.019)
Prerequisites are sufficient	0.014	0.025	-0.000	-0.000
	(0.009)	(0.020)	(0.012)	(0.013)
Workload is consistent with the ECTS	0.121***	0.112***	0.131***	0.070***
	(0.013)	(0.024)	(0.014)	(0.011)
Your interest for the subject	0.167***	0.072**	0.151***	0.138***
	(0.015)	(0.028)	(0.018)	(0.017)
Constant	-0.074	-0.109	0.099	0.003
	(0.153)	(0.025)	(0.155)	(0.144)
R-squared	0.788	0.817	0.728	0.827
Observations	1,641	487	1,574	1,160

Note: OLS estimates. Standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1. All items are evaluated on a scale ranging from 1 to 10. Coefficients in the table represent the effect on overall satisfaction of increasing by one point the evaluation of the items in column one.

Table 3. Decomposition of the variance of SET for the vignette courses (percentages). By major.

	Variance between courses	Variance within courses		
	% of total variance	% of total variance	% of (2a) due to noise	% of (2a) due to reporting heterogeneity
	(1)	(2a)	(2b)	(2c)
Economics	0.287	0.713	0.653	0.347
Engineering	0.323	0.677	0.538	0.462
Law	0.193	0.807	0.743	0.267
Medicine	0.204	0.796	0.750	0.250

Table 4. Tests for sorting. By major.

	Economics	Engineering	Law	Medicine
	(1)	(2)	(3)	(4)
θ_0	0.188*	0.340***	-0.561**	-0.149
	(0.095)	(0.095)	(0.187)	(0.168)
θ_1	0.980***	0.967***	1.075***	1.020***
	(0.012)	(0.012)	(0.023)	(0.022)
Observations	492	144	376	436
R-squared	0.933	0.978	0.853	0.835
P-values for:				
H0: $\theta_0 = 0$	0.048	<0.001	0.003	0.377
H0: $\theta_1 = 1$	0.092	0.008	0.001	0.369
H0: $(\theta_1 = 0; \theta_1 = 1)$	0.002	<0.001	<0.001	0.664

Note: OLS estimation of the linear model $y_{.j|e} = \theta_0 + \theta_1 y_{.j} + \mu_{ej}$. For each vignette j, $y_{.j|e}$ is the average evaluation of vignette j provided by the students choosing the elective *e*, *e=1,…,*E_j, while $y_{.j}$ is the average evaluation provided by all students in the stratum which vignette j belongs to. The null hypothesis is θ_0=0 and θ_1=1. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Figures

Figure 1. Average evaluation of vignettes with 95% confidence intervals. By major.

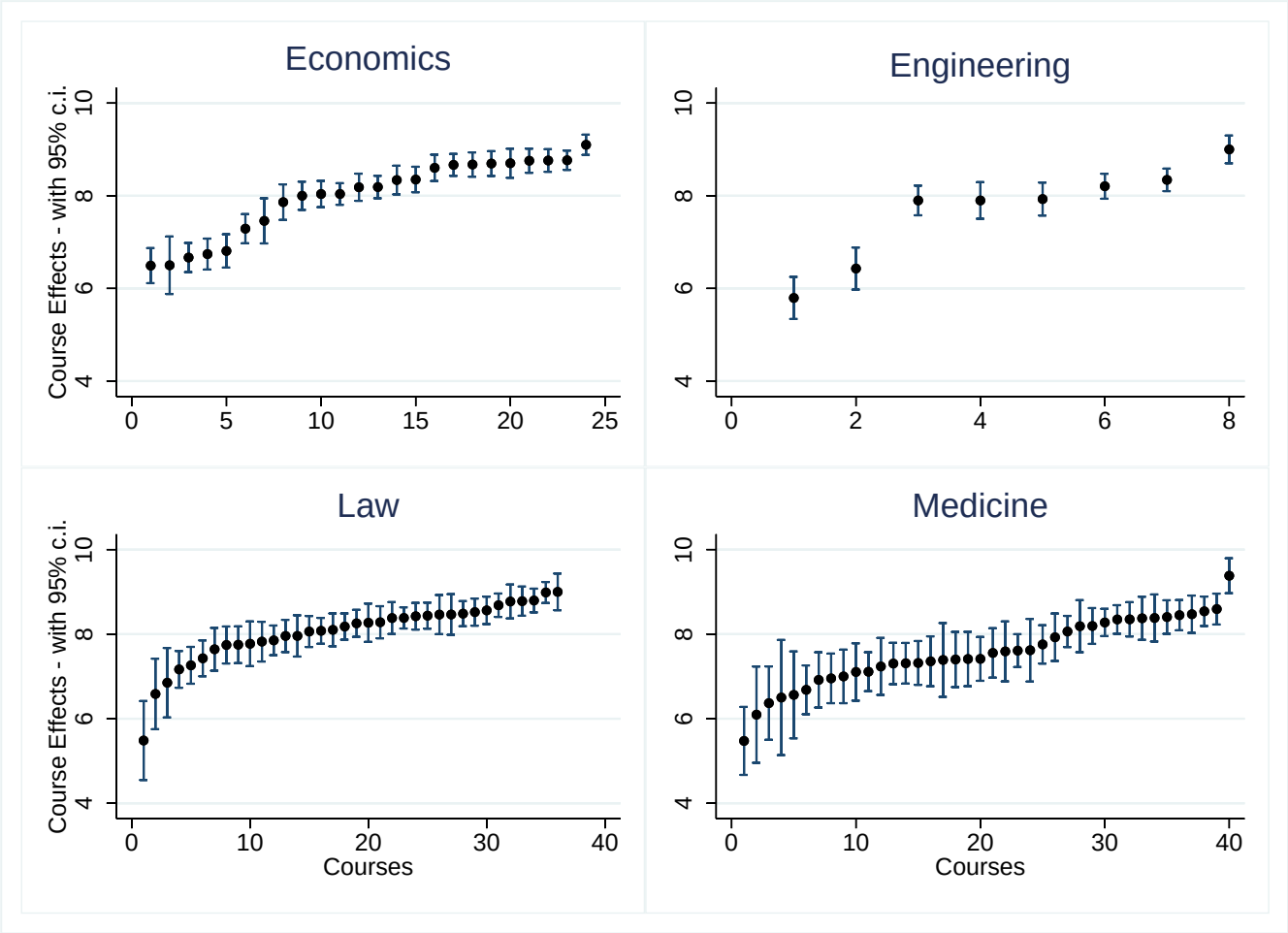
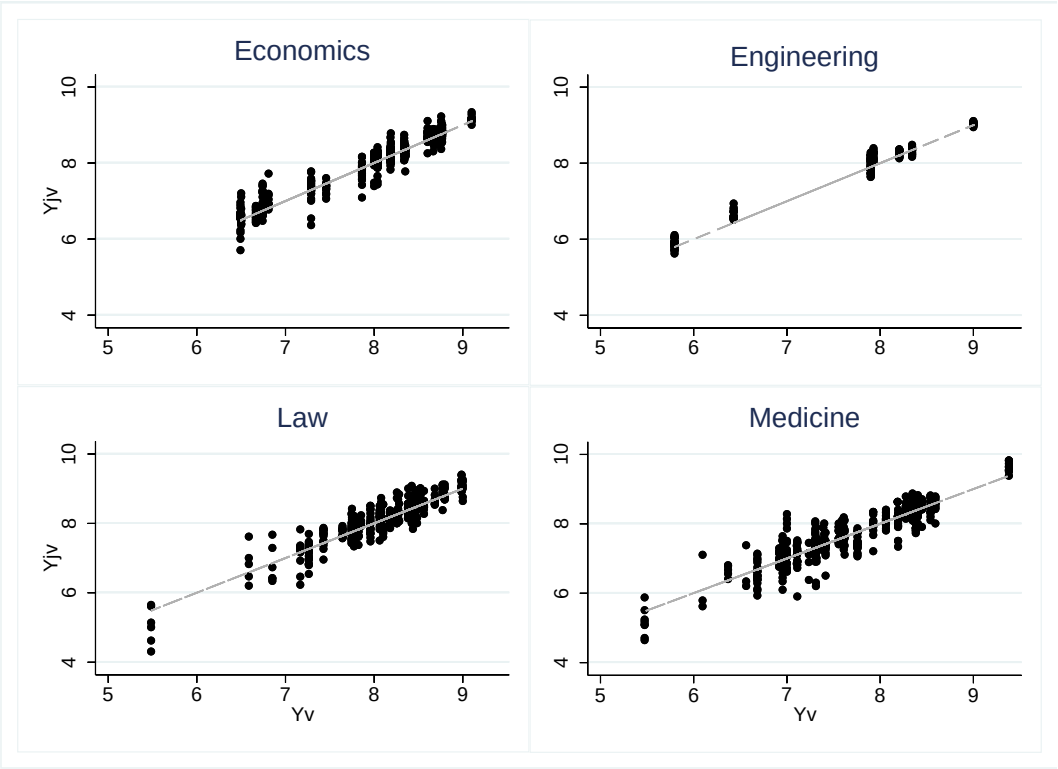
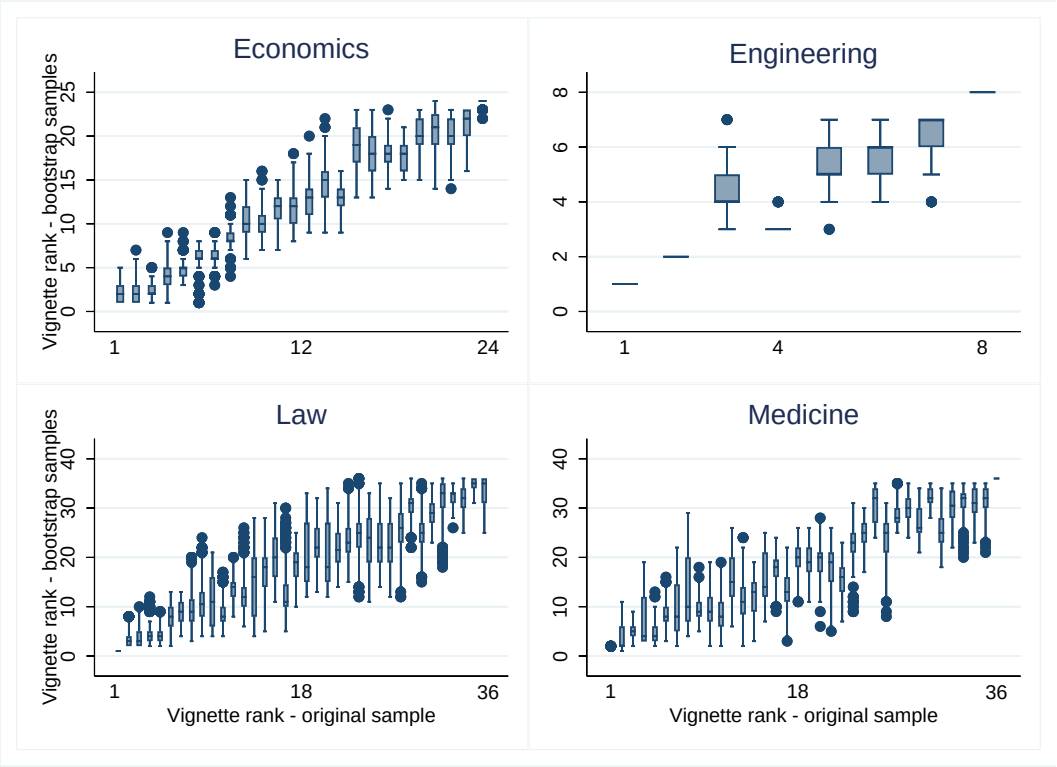


Figure 2. Average evaluation of vignettes by students choosing elective *e*, *e=1,…, E* vs overall average evaluation of vignettes – all strata pooled. By major.



Note: For each vignette *j*, the horizontal axis reports $y_{.j}$, the average evaluation provided by the students in the stratum which vignette *j* belongs to; the vertical axis reports $y_{.j|e}$, the average evaluation of vignette *j* provided by the students choosing elective *e*, *e=1,…,*E_j.

Figure 3. Boxplots of bootstrapped rankings of courses. By major.



Note: 200 replications. In each replication, we randomly draw one elective course e per stratum, compute $y_{.j|e}$ for all vignettes which belong to the stratum, pool all strata of the major and define the corresponding rank of each vignette. For each vignette the graph reports the boxplot of the distribution of the rank positions occupied by the vignette across the replications.

Appendix

Table A1 – Derivation of the study sample

	Economics			Engineering			Law			Medicine		
	Students	Courses	Strata	Students	Courses	Strata	Students	Courses	Strata	Students	Courses	Strata
	(1a)	(1b)	(1c)	(2a)	(2b)	(2c)	(3a)	(3b)	(3c)	(4a)	(4b)	(4c)
1. *Reference population*: at least one evaluation as attendee	598	201	6	242	79	3	1317	210	9	953	987	12
2. Keep only students with at least 3 evaluations	561	201	6	232	79	3	944	204	9	841	981	12
						Vignette definition at this stage						
3. Keep only students who evaluated at least 3 vignettes.	465	201	6	201	79	3	544	204	9	492	981	12
4. Keep only students with variation in their vignette evaluations	443	201	6	195	79	3	477	204	9	457	981	12
5. Keep only strata with variation in average vignette evaluations	443	201	6	195	79	3	477	204	9	405	927	11
6. Keep only strata with no selection issues w.r.t. average vignette evaluations between students who evaluate at least one vignette in 2. and 5.	443	201	6	133	46	2	477	204	9	339	775	10
7. *Final sample*: keep only electives evaluated by at least 10 students	443	147	6	133	44	2	477	130	9	339	149	10

Table A2 – Observable characteristics in the study sample and the reference population

	Number of students		Female		Local-born student		Year of birth (19-)		High school grade (60-100)	
	Reference population	Final sample	Reference population	Final sample	Reference population	Final sample	Reference population	Final sample	Reference population	Final sample
	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)	(4a)	(4b)	(5a)	(5b)
Economics	598	443	0.56	0.60	0.77	0.77	92.82	92.89	94.35	94.76
Engineering	242	133	0.46	0.53	0.86	0.83	92.62	93.30	82.80	82.02
Law	1317	477	0.63	0.66	0.83	0.86	92.46	92.66	79.70	82.34
Medicine	953	339	0.51	0.50	0.73	0.74	92.64	92.85	91.23	92.54

Table A3 Description of the final sample – Economics

	Pooled	Stratum 1	Stratum 2	Stratum 3	Stratum 4	Stratum 5	Stratum 6
Number of students	443	57	84	68	91	53	90
Number of courses							
Vignettes	24	4	4	4	4	4	4
Electives	123	23	27	22	26	13	12
Evaluations by student							
Vignettes	3.77	3.82	3.65	3.76	3.80	3.72	3.83
Electives	10.39	11.68	12.14	11.09	11.14	9.32	7.3
Evaluations by course							
Vignettes	69.54	54.5	76.75	64	86.5	49.25	86.25
Electives	37.44	28.96	37.78	34.27	39	38	54.75
Coverage							
Vignettes - at definition	0.86	0.92	0.79	0.88	0.83	0.86	0.88
Vignettes - in final sample	0.94	0.96	0.91	0.94	0.95	0.93	0.96
Electives – in final sample	0.51	0.51	0.45	0.50	0.43	0.72	0.61

Table A4 Description of the final sample – Engineering

	Pooled	Stratum 1	Stratum 2
Number of students	133	74	59
Number of courses			
Vignettes	8	4	4
Electives	36	22	14
Evaluations by student			
Vignettes	3.84	3.89	3.78
Electives	13.44	16.07	10.14
Evaluations by course			
Vignettes	63.88	72	55.75
Electives	49.64	54.05	42.71
Coverage			
Vignettes - at definition	0.91	0.93	0.88
Vignettes - in final sample	0.96	0.97	0.94
Electives – in final sample	0.73	0.73	0.72

Table A5 Description of the final sample – Law

	Pooled	Stratum 1	Stratum 2	Stratum 3	Stratum 4	Stratum 5	Stratum 6	Stratum 7	Stratum 8	Stratum 9
Number of students	477	52	57	33	56	78	62	44	51	44
Number of courses										
Vignettes	36	4	4	4	4	4	4	4	4	4
Electives	94	10	11	6	16	14	16	5	8	8
Evaluations by student										
Vignettes	3.48	3.42	3.51	3.39	3.48	3.54	3.65	3.27	3.49	3.5
Electives	4.16	2.98	4.30	2.48	5.93	4.95	6.15	2.18	3.19	3.23
Evaluations by course										
Vignettes	46.17	44.5	50	27.75	48.75	69	56.5	36	44.5	38.5
Electives	21.09	15.5	22.27	13.67	20.75	27.57	23.81	19.2	20.38	17.75
Coverage										
Vignettes - at definition	0.67	0.68	0.68	0.60	0.64	0.75	0.70	0.62	0.73	0.66
Vignettes - in final sample	0.87	0.86	0.88	0.84	0.87	0.88	0.91	0.82	0.87	0.88
Electives – in final sample	0.38	0.30	0.39	0.41	0.37	0.35	0.38	0.44	0.40	0.40

Table A6 Description of the final sample – Medicine

	Pooled	Stratum 1	Stratum 2	Stratum 3	Stratum 4	Stratum 5	Stratum 6	Stratum 7	Stratum 8	Stratum 9	Stratum 10
Number of students	339	22	30	25	51	21	18	22	55	52	43
Number of courses											
Vignettes	40	4	4	4	4	4	4	4	4	4	4
Electives	109	8	12	9	17	5	0	3	17	17	21
Evaluations by student											
Vignettes	3.55	3.73	3.77	3.64	3.55	3.33	3.44	3.45	3.55	3.55	3.53
Electives	6.01	4.73	6.2	5.36	6.39	2.71	-	1.77	7.58	6.65	9.98
Evaluations by course											
Vignettes	30.1	20.5	27.5	22.75	45.25	17.5	15.5	19	48.75	46.25	38
Electives	18.7	13	15.5	14.89	19.18	11.4	-	13	24.53	20.35	20.43
Coverage											
Vignettes - at definition	0.66	0.7	0.68	0.67	0.67	0.61	0.59	0.60	0.70	0.75	0.64
Vignettes - in final sample	0.89	0.93	0.92	0.91	0.89	0.83	0.86	0.86	0.89	0.89	0.88
Electives – in final sample	0.47	0.59	0.52	0.60	0.38	0.54	-	0.59	0.45	0.39	0.48

Figure A1. Average evaluation of vignettes in the reference population vs final sample – including also dropped strata. By major.

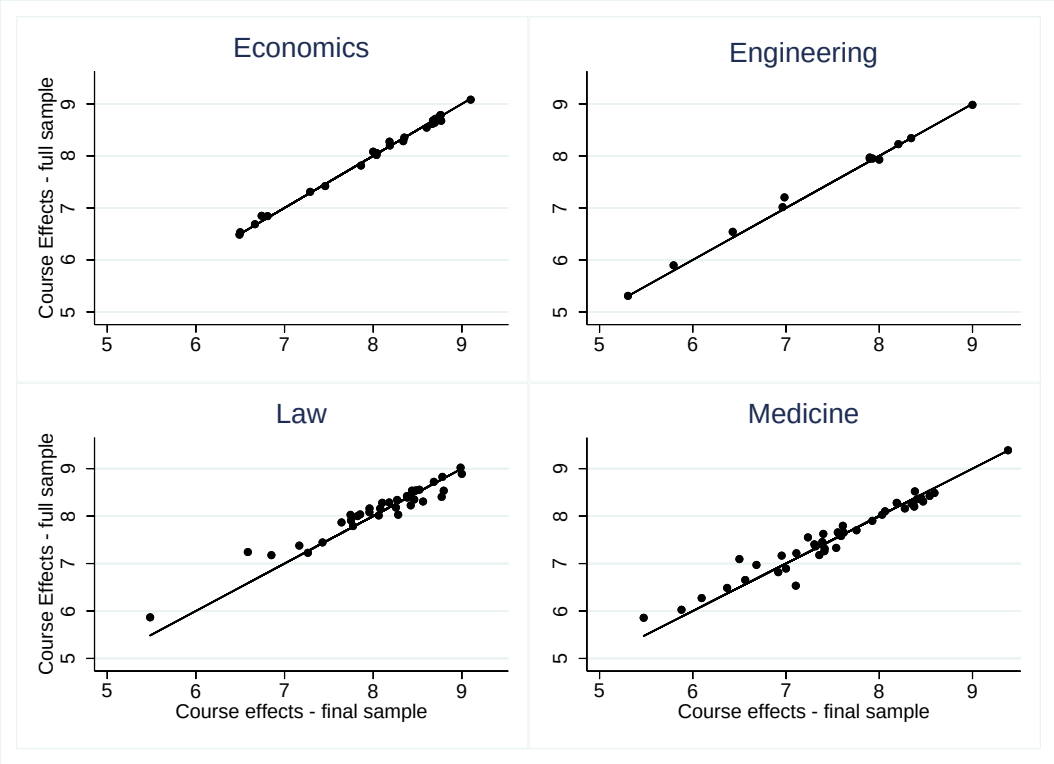


Figure A2. Sampling over total variance as course size increases (see main text). By major.

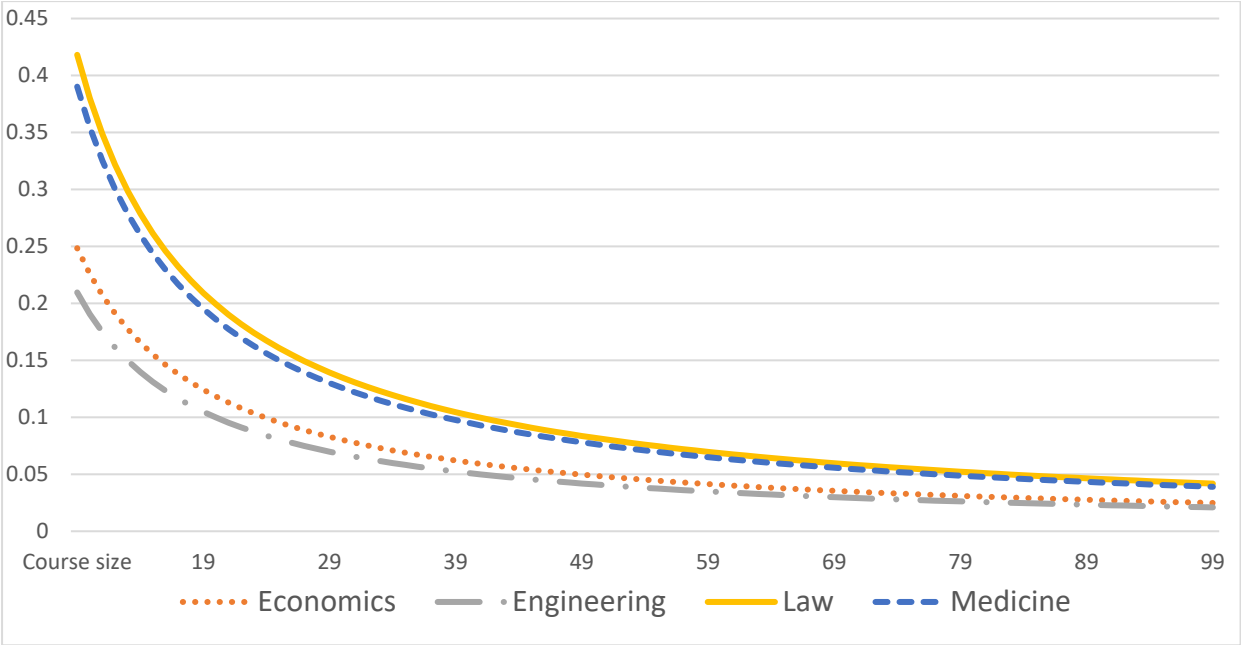
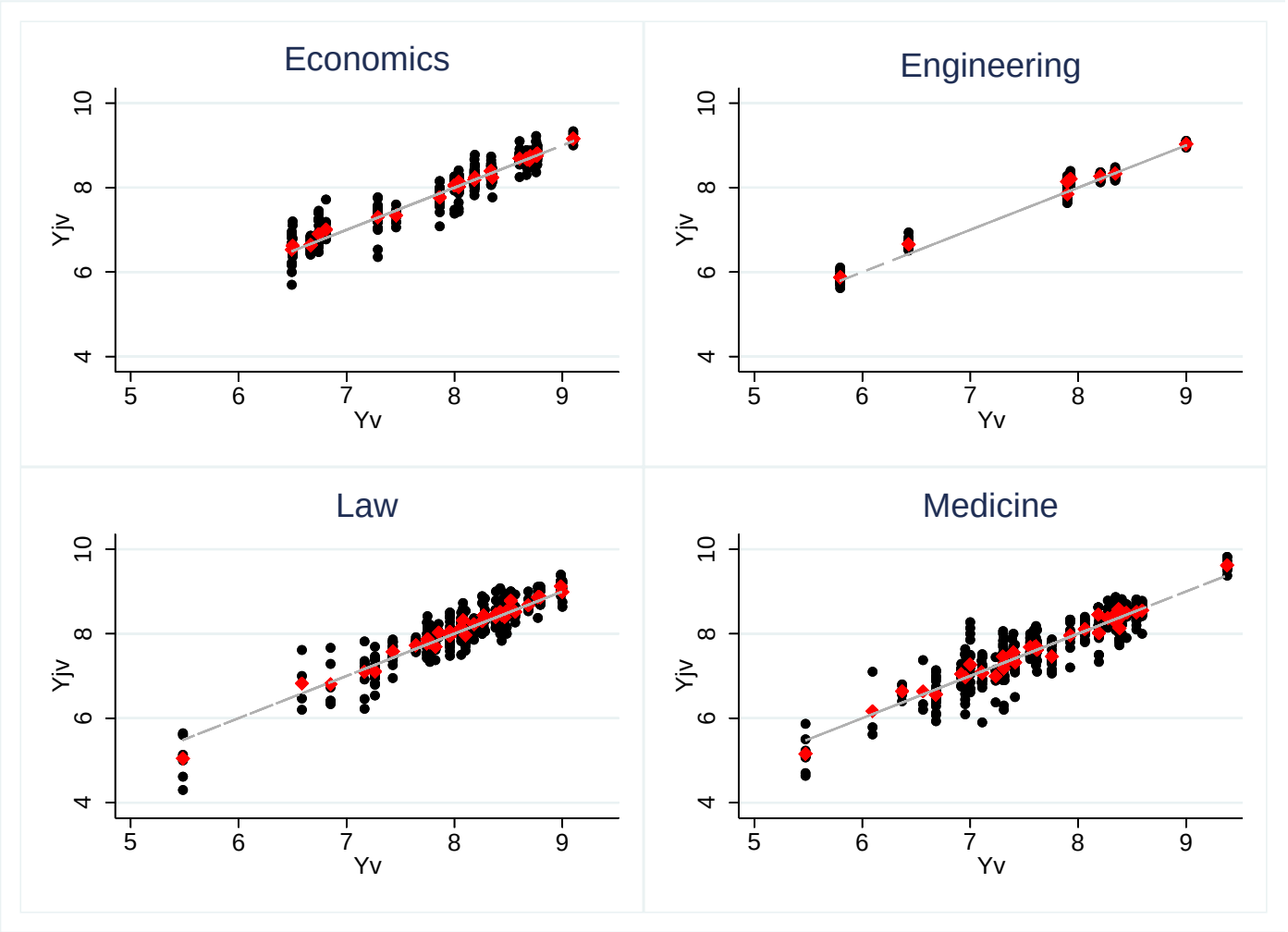
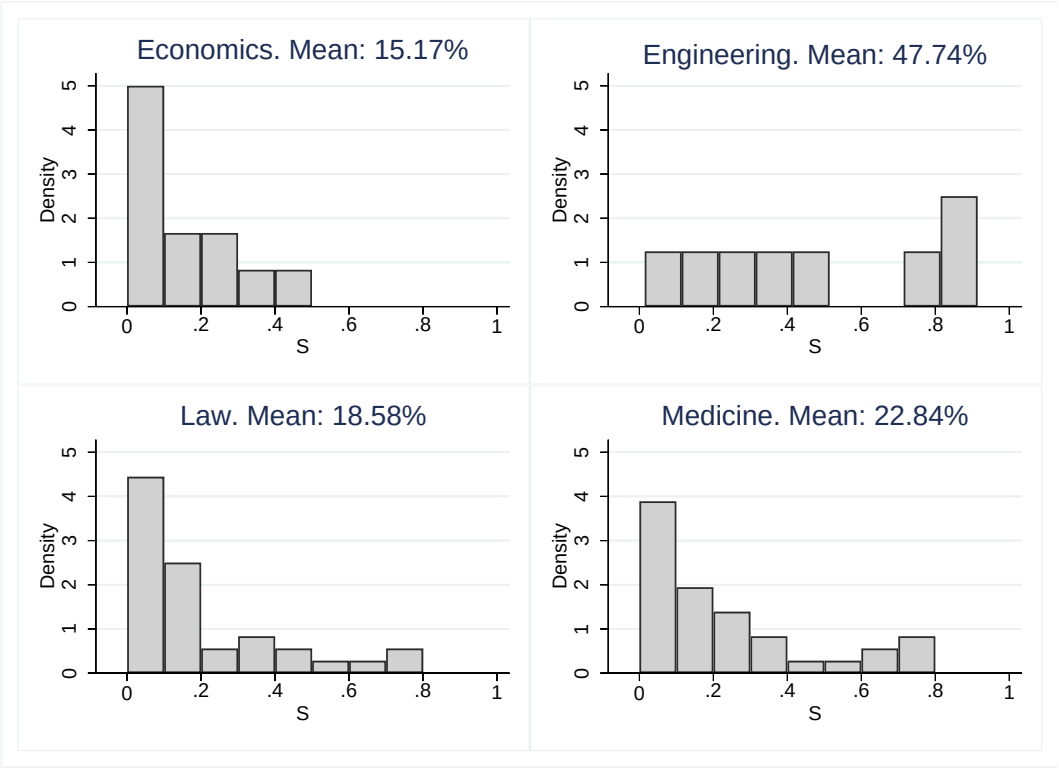


Figure A3. Dispersion of $y_{.j|e}$ and the average of $y_{.j|e}$ for all vignettes (see main text). By major.



Note: red dots correspond to the average of $y_{.j|e}$, by vignette.

Figure A4. Distribution of ratio S, the importance of sorting in vignette evaluation, across vignettes (see main text). By major.



Note: histogram bins have width 0.1.