Labour Economics 89 (2024) 102567

Contents lists available at ScienceDirect

Labour Economics

journal homepage: www.elsevier.com/locate/labeco

If (my) 6 was (your) 9. Reporting heterogeneity in student evaluations of teaching☆

Marco Bertoni [a], Enrico Rettore [b,*], Lorenzo Rocco [c]

[a] University of Padova, HEDG and IZA, Italy
[b] University of Padova, FBK-IRVAPP and IZA, Italy
[c] University of Padova and IZA, Italy

ARTICLE INFO

JEL classification:
I23
I28
D63

Keywords:
Student evaluations of teaching
Reporting heterogeneity
Selection

ABSTRACT

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 enable us to track all evaluations produced by each student, we 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.

1. 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 now 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. Typically, individual SET are averaged by course or class, and average SET is the one-dimensional indicator that professors and deans look at.

In addition to professors and deans, there is evidence that SET are relevant also to prospective students. In a recent paper, using data from the universe of Dutch universities, de Koning et al. (2022) study how high school graduates choose their university/program. They find that '…student satisfaction scores matter for enrolment. An increase in a program's student satisfaction score leads to higher levels of enrolment, whereas an increase in the student satisfaction scores of [program's] substitutes leads to lower levels of enrolment…'. That is, according to these results, SET are used by students not only to compare courses within major but also to compare programs both within and across universities.

Given the increasing stakes which depend on SET, it is not surprising that the validity of SET has been put under scrutiny by scholars. A number of studies have concluded that average 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. Such negative conclusions would turn even more radical if SET were considered as purely ordinal evaluations, lacking a cardinal meaning. For instance, if a course rated 10 was preferred to – but not twice as good as – a course rated 5, then average course ratings from different students would make little sense as measures of teaching quality.

Notwithstanding these concerns, SET scores are widely used by schools to evaluate their teachers, often in tournament-like

☆ Marco Bertoni: Department of Economics and Management "Marco Fanno", University of Padova. Via del Santo 33, 35123 Padova - Italy. Email: marco.bertoni@unipd.it. Enrico Rettore (corresponding author): Department of Economics and Management "Marco Fanno", University of Padova. Via del Santo 33, 35123 Padova - Italy. Email: enrico.rettore@unipd.it. Lorenzo Rocco: Department of Economics and Management "Marco Fanno", University of Padova. Via del Santo 33, 35123 Padova - Italy. Email: lorenzo.rocco@unipd.it. Acknowledgments and disclaimers are reported at the end of the main text.

* Corresponding author.
E-mail address: enrico.rettore@unipd.it (E. Rettore).

https://doi.org/10.1016/j.labeco.2024.102567
Received 8 November 2023; Received in revised form 2 May 2024; Accepted 6 May 2024
Available online 17 May 2024

comparisons. In this paper, we show that even if SET were not manipulable, captured only teaching effectiveness, and were cardinal – in the following we refer to this as the best-case scenario – schools adopting SET would still face an additional and so far overlooked problem: heterogeneity in students' reporting styles.

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 the same course as "fair" might rate it differently if the former thinks that a grade of 6 out of 10 corresponds with a "fair" evaluation, while, according to the latter, 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 5 and 8, while the other grades between 1 and 10.

Reporting heterogeneity may have pervasive implications for the comparability of courses evaluated by different students. If students are randomly distributed across courses, or if all courses are attended by the same students, then the distribution of students' reporting styles will be identical in all courses and reporting heterogeneity will not bias the relative evaluation of the courses.[1] Instead, if students with a more lenient reporting style sort in some courses, while students with a more strict reporting style concentrate in others, then reporting heterogeneity prevents comparability and course average SET cannot be used to portray a valid ranking of course quality.

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 document that a sizeable proportion of the overall variability in individual SET is due to reporting heterogeneity. Furthermore, we test and reject the hypothesis of random sorting in favour of the hypothesis of sorting on reporting style and we show that the rank of courses by their average SET is affected by sorting on reporting style to a non-negligible extent.

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 both 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. Within strata defined by the intersection of major, cohort and track, the courses which are attended and evaluated by the large majority of students, a condition which approximates complete coverage, play the role of vignettes. With a slight abuse of language, we refer to the remaining courses as electives.

Since vignette courses are attended and evaluated by approximately all students of a stratum, in the best-case scenario 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: *i)* variation due to systematic differences between-courses; *ii)* within-course variation due to student-specific reporting styles; and *iii)* 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.

An alternative interpretation of the variation in ratings within courses can be in terms of student-by-course match effects, such as student specific interest in the subject being taught. For example, the average rating of an introductory mathematics course would depend on the aversion to mathematics in the student population, regardless of the quality of the mathematics course offered. As a result, comparing courses in different subjects (e.g. mathematics and accounting) would confound aspects of quality with student tastes, making it impossible to draw conclusions about teaching quality alone - even in the absence of reporting heterogeneity and sorting.[2] Several pieces of evidence help us to show that this is not a relevant concern in our setting. First, we find that students' self-reported interest in the subject taught can explain only about 10 % of the variation in the reported overall course quality of the vignettes, while more than 40 % of this variation can instead be attributed to students' perception of the teacher's ability to motivate students, explain things clearly, and select adequate teaching material. Second, if tastes were the main driver of reported overall quality, we would expect electives – which students choose to match their interests – to be rated higher than vignettes. Instead, we find that vignettes consistently receive higher ratings than electives across all majors. Finally, we show that all of our main results are unaffected when we remove the component of overall satisfaction that is due to students' interest in the subject being taught – the closest proxy for a matching effect that we can observe in our data.

Assuming match effects being negligible, we test whether students' sorting across elective 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 subgroups 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 chosen subset of evaluators. Also, the top-ranked vignette can lose up to 10 positions out of 36 and turn to be a mid-rank course.

Overall, our results suggest that the combination of reporting heterogeneity and sorting – on top of sampling variability – largely limits the comparability of average evaluations of courses. Accordingly, schools should be cautious in relying on SET to incentivise, promote or

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

[2] Linask and Monsk, 2018, highlight that students may sort across elective courses depending on their interest for the subject taught. In their empirical analysis, they show that controlling for students' prior interest in the subject taught substantially alters the ranking of elective courses.

[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).

hire teachers, especially within tournament-like schemes. Or, at the very least, they should aim at reducing sampling variability by keeping the size of the courses relatively large[4] and offer sufficiently many vignette courses to estimate students' reporting styles and correct SET in elective courses.

Finally, we argue that correcting SET bias by exploiting the anchoring provided by vignettes is not necessarily beneficial. In our data, for roughly two thirds of the electives the mean squared error of the corrected average SET of elective courses exceeds the one associated with the raw average SET. By requiring the additional step of estimating elective-specific corrections, the bias due to reporting heterogeneity is reduced at the price of increasing the sampling variability of the estimated course effects.

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 review

A vast literature has analysed SET and debated on their reliability and validity. In theory, SET aim to measure teaching effectiveness, a concept that is inherently difficult to define. In practice, they serve many purposes, including feedback to help teachers improve their courses, and evaluate teaching in promotion, rewarding or hiring procedures.

In this Section we review some relevant seminal papers. We 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 also used follow-on courses as indicators of teaching effectiveness under the assumption 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 evaluate more positively those professors who contribute less to their performances in 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, teachers, 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. MacNell et al. (2015), Boring (2017) and Mengel et al. (2019) exploit settings where instructors (and hence their gender) are randomly assigned to students and confirm this finding.[5]

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 subjective scale to express their appreciation for a course.[6] 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 5 s.

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 Wolbring and Treischl (2015) 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 adding a further concern, that of reporting heterogeneity, 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 degree (*laurea triennale*) in Economics; a bachelor degree in

[4] Provided, of course, that a increasing class size is not detrimental to teaching quality.

[5] 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.

[6] Additional negative implications of the use of ordinal measures are discussed in Bond and Lang, 2013, 2019.

Civil Engineering and its natural continuation, the master degree (*laurea magistrale*) in Civil Engineering; the five-year degree (*laurea a ciclo unico*) in Law and the six-year degree in Medicine and Surgery. Below, we treat the bachelor and the master degree in Civil Engineering as a unique five-year degree since the large majority of bachelor students continue to the master.

To hold class size manageable, students in each major are further split in tracks, defined according to the initial letter of students' family name, to the location of teaching rooms, 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 training at the hospital.

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. Given the rules generating the tracks are constant across cohorts and the fact that the three matriculation cohorts are adjacent, we consider strata as representative of a major's population and pool them in our analysis.

We define a course as a learning unit taught by one professor to students belonging to a given stratum.[7] 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, and we consider the sub-unit taught by each teacher as an autonomous course. We also treat a given learning unit taught by the same professor in two academic years as two separate courses because, on the one hand, each year course content of teaching methods could vary to some extent and, on the other hand, this choice is consistent with the policies of the University that we study, where data on average SETs by course and teacher are published on a yearly basis – without pooling new data with existing ones – and the judgement of teaching quality is carried out every academic year.

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.

Among these, we define as vignettes the four courses with the highest coverage in each stratum, where coverage is defined as the ratio between the total number of evaluations turned in for a course and the total number of students in the stratum. We refer to electives to indicate all courses which are not vignettes. By construction each stratum maps to some vignettes and a vignette maps to a unique stratum.

The key feature of our dataset is that we can track all the evaluations provided by a specific student. This is possible because students fill in SET questionnaires when they register for exams from their electronic personal account on the University's web system, which records their identity. However, students' identity is not transmitted to teachers, who only receive aggregate data on their evaluations. We have been granted access to anonymised micro-level data for the specific purposes of this project.

[7] 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.

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

	Economics (1)	Engineering (2)	Law (3)	Medicine (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

3.2. Sample selection

Starting from the full sample of evaluations provided by students in our data, we apply several selection criteria dictated by the need to reach 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, and includes 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. We illustrate the details of our sample selection procedure in Appendix A, where we also provide evidence that the selected sample, although smaller in size, is decently representative of the population.

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 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 % 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 coherent with the number of credits
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 on a discrete scale ranging from 1-lowest to 10-highest.[9] Following most of the literature, we take students' overall satisfaction as the main indicator of SET. Importantly, the university administration also focuses on overall satisfaction in its official reports.

The average overall satisfaction for the vignettes of each major (pooling all strata), along with the corresponding 95 % confidence interval, are displayed in Fig. 1. By construction, in our best-case scenario the average SET for vignettes do not suffer of reporting bias and provide comparable estimates of course quality. We notice that they are rather compressed, and it is possible to statistically distinguish only between top and poor performers. Since vignettes are the courses evaluated by the largest number of students, this result casts concerns about the reliability of the ranks for smaller courses.

Before moving to our analysis on sorting and reporting heterogeneity, we assess the extent to which overall satisfaction is a comprehensive measure of satisfaction with the various aspects of the course rated by students, described above. We do so by regressing overall satisfaction on satisfaction with each item, separately for each major. Table 2 reports the estimated regression coefficients, the overall R-squared, and the Shapley-Owen decomposition of the R-squared, that assess the explanatory power of each individual regressor (Hüttner and Sunder, 2012). We find that the included items explain a large share of the variation in overall satisfaction, as the R-squared is always above 0.7. The items more related with overall satisfaction, and that jointly explain more than 40 % of its variation, are those capturing the abilities of the instructor to motivate the class, to teach in a clear way, and to select and prepare high-quality teaching material. Other dimensions, including workload adequacy and the organizational dimensions of the course, are less relevant in determining students' satisfaction. Importantly, "interest for the subject taught" – the closest proxy of match effects that is available in our data – only explains up to 11 % of the variation of overall satisfaction.[10]

Evidence that student specific tastes for a course are of little relevance in the determination of course satisfaction also comes from the comparison of each student's average evaluations of electives and vignette courses. If prior taste was the main driver of satisfaction, we would expect that elective courses – that students likely choose on the basis of their interests for the subject taught – receive higher evaluations than vignettes. Instead, we find that – consistently across all majors – vignettes receive higher evaluations than elective courses, even after including student-by-semester fixed effects to account for selection in elective courses and learning effects.[11]

We further show that match effects are not the primary driver of our findings by replicating all our main empirical results after purging "overall satisfaction" from its component related with "personal interest", to eliminate taste heterogeneity unrelated to teaching quality.[12] We anticipate that the results do not vary significantly.[13]

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 residual one. 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 $x_{i\cdot}$ the (sample) average of x_{ij} for student i across courses, by $x_{\cdot j}$ the (sample) average of x_{ij} for course j across students, and by $x_{\cdot\cdot}$ the overall (sample) average of x_{ij}.

We model student i 's evaluation of course j, denoted y_{ij}, as

$$y_{ij} = \alpha_i + \beta_i \gamma_j + \varepsilon_{ij} \tag{1}$$

where γ_j is a course specific component, and ε_{ij} is an individual-by-course component, with $\varepsilon_{\cdot j}$ converging to zero as the number of students evaluating course j gets larger. Each student reports γ_j using the student-specific linear transformation (1), which includes a noise component, ε_{ij}. The parameter α_i captures how lenient a student is when he or she rates a course, while the parameter β_i captures his/her sensitivity to variations in γ_j.[14]

We assume that ε_{ij} is uncorrelated with both the course specific component γ_j and the student reporting style and it is independently distributed across students and courses.

Assumption 1. In the population, $E(\varepsilon_{ij}) = 0$; $cov\left(\varepsilon_{ij}, \gamma_j\right) = 0$; $cov(\varepsilon_{ij}, \alpha_i) = 0$; $cov(\varepsilon_{ij}, \beta_i) = 0$; $cov(\varepsilon_{ij}, \varepsilon_{ij'}) = 0$ for $j \neq j'$ and $cov(\varepsilon_{ij}, \varepsilon_{i'j}) = 0$ for $i \neq i'$.

Up to the noise component ε_{ij}, this model formalizes the intuition in King et al. (2004). Individual evaluations of course j differ due to two individual specific parameters shifting and stretching, respectively, the scale of measurement. Knowledge of those individual specific parameters would allow to make those evaluations comparable by shifting and rescaling the reported y_{ij} through the mapping:

$$\left(y_{ij} - \alpha_i\right) \Big/ \beta_i$$

(see their Fig. 1 and the discussion thereof). We also posit:

Assumption 2. Within each major, the averages of α_i and β_i across all students are 0 and 1 respectively.

Assumption 2 is a normalization that comes as a straightforward consequence of the impossibility to compare SET across majors due to the lack of any common vignette. Assumption 2 implies that for vignette courses $\alpha_{\cdot} = 0$ and $\beta_{\cdot} = 1$, while this is not necessarily true for elective

[8] The question students face reads as follows: "Overall, how satisfied are you with this course?"

[9] Rivera and Tilcsik, 2019, use a natural experiment to show that the scale of the evaluations (a 6-point or a 10-point one) affects the gender gap in SETs in male-dominated fields. In our setup, the response scale is common across questions, years, and majors.

[10] These conclusions are unaltered when we include students' and teachers' observable characteristics as additional controls.

[11] The coefficients on a "vignette course" dummy in a regression of overall satisfaction on this dummy and student-by-semester fixed effects is equal to 1.09 points for economics, 1.28 points for engineering, 0.82 points for law, and 0.29 points for medicine. All estimates are significant at better than the 1 percent level of confidence.

[12] We achieve this by subtracting from overall satisfaction the coefficient of "interest for the subject taught" from the regressions reported in Table 2, multiplied by the value of "interest for the subject taught" reported by each student.

[13] We have also used principal component analysis to extract a single factor out of all SET questions present in the questionnaire, and replicated our main analyses using this factor. The results – not reported to save space, but available from the authors – are wholly comparable to the ones reported in the paper.

[14] This specification meets the requirements of response consistency and vignette equivalence discussed in King et al., 2004. We treat γ_j as cardinal, so that each affine transformation yields cardinal scores. Cardinality is necessary to define and decompose the variance of y_{ij}.

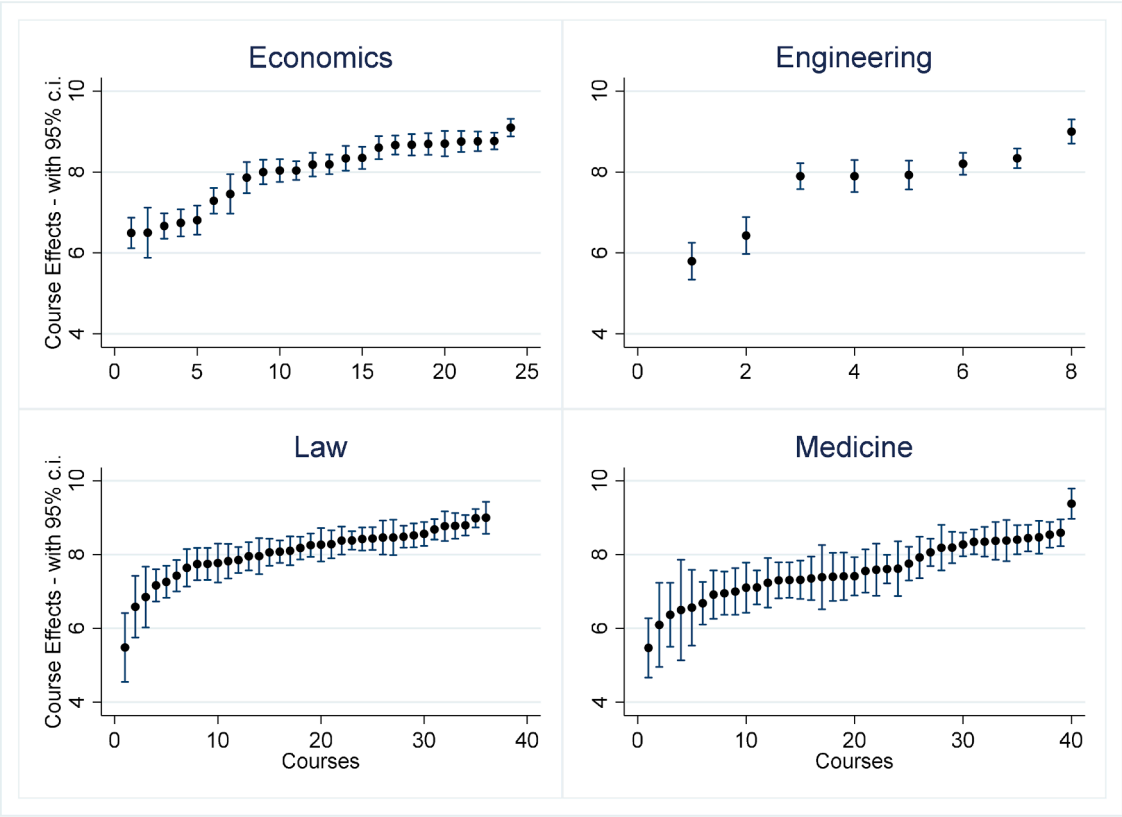

Fig. 1. Average evaluation of vignettes with 95 % confidence intervals. By major.

courses. The average evaluation of vignette j is thus $y_{\cdot j} = \gamma_j + \varepsilon_{\cdot j}$. Moreover, given that vignettes are attended by many students, Assumption 1 implies that $\varepsilon_{\cdot j}$ can be approximated to 0, so that $y_{\cdot j} = \gamma_j$. In other words, in vignettes 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 evaluations which is common across all students belonging to the same stratum and corresponds to their average evaluation of teacher quality for course j. Reporting heterogeneity arises by allowing (α_i, β_i) to vary across students. Each student's reporting function is a linear transformation of the course component γ_j plus a zero-mean residual. The latter includes trembling-hand errors in evaluation and random shocks. For simplicity, we refer to the ε_{ij} component of student evaluations as noise, to highlight its unsystematic nature.

4.1. Variance decomposition for vignette courses

Now, let us focus on vignettes. In a major there are N students and K vignettes. Students are indexed by $i = 1, \ldots, N$ and vignettes by $j = 1, \ldots, K$. Each vignette j is attended by n_j students and each student i evaluates k_i vignettes.

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

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

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 $y_{\cdot j} = \gamma_j$, the between-course variability is unaffected by students' reporting style and reflects only genuine differences between courses. The within-course deviance accounts for the variability among individual evaluations and combines reporting heterogeneity and all residual variability (noise). Thanks to the panel nature of our data, we can take these two factors apart.

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

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

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

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

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

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 requires us to estimate ε_{ij}. To this end, we estimate model (1) on vignette evaluations, separately by major. Recalling that for each vignette j the component γ_j (approximately) corresponds to $y_{\cdot j}$, we regress y_{ij} on a full set of individual dummies and a full set of interactions between γ_j and individual dummies. The parameters of this model identify the vector of individual-specific intercept sand slopes α_i

Table 2
OLS estimates of the association between overall satisfaction and satisfaction with other course aspects.

	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)
Contribution to the R-squared	[0.082]	[0.073]	[0.088]	[0.086]
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)
Contribution to the R-squared	[0.042]	[0.039]	[0.042]	[0.044]
Punctuality of the instructor	−0.003	0.053**	0.027	0.042***
	(0.015)	(0.026)	(0.017)	(0.016)
Contribution to the R-squared	[0.018]	[0.061]	[0.030]	[0.040]
Quality of lecture notes/ reference books	0.078***	0.099***	0.056***	0.069***
	(0.013)	(0.021)	(0.017)	(0.016)
Contribution to the R-squared	[0.094]	[0.060]	[0.071]	[0.088]
Instructor is able to motivate the class	0.213***	0.195***	0.228***	0.319***
	(0.017)	(0.034)	(0.019)	(0.019)
Contribution to the R-squared	[0.165]	[0.172]	[0.158]	[0.215]
Instructor teaches in a clear way	0.284***	0.342***	0.275***	0.252***
	(0.017)	(0.032)	(0.022)	(0.019)
Contribution to the R-squared	[0.190]	[0.206]	[0.159]	[0.182]
Prerequisites are sufficient	0.014	0.025	−0.000	−0.000
	(0.009)	(0.020)	(0.012)	(0.013)
Contribution to the R-squared	[0.040]	[0.053]	[0.022]	[0.027]
Workload is coherent with the number of credits	0.121***	0.112***	0.131***	0.070***
	(0.013)	(0.024)	(0.014)	(0.011)
Contribution to the R-squared	[0.066]	[0.076]	[0.062]	[0.029]
Your interest for the subject	0.167***	0.072**	0.151***	0.138***
	(0.015)	(0.028)	(0.018)	(0.017)
Contribution to the R-squared	[0.091]	[0.077]	[0.095]	[0.116]
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	1641	487	1574	1160

Note: OLS estimates. Standard errors in parentheses. R-squared decomposition values, reported in square brackets, sum to the overall R-squared.
*** $p < 0.01$,.
** $p < 0.05$, *$p < 0.1$. 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 1. Observations with missing evaluations for any item in the questionnaire are dropped.

and β_i in model (1),[15] and we use their estimates to obtain the residuals $\widehat{\varepsilon}_{ij} = y_{ij} - \widehat{\alpha}_i - \widehat{\beta}_i\gamma_j$. Finally, the estimated variance of residuals is:

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

where we take into account that two degrees of freedom per student are lost in the estimation of $\widehat{\alpha}_i$ and $\widehat{\beta}_i$. Expression (5) is an unbiased estimator of the variance of the noise component in (4).

Table 3 reports the results of this decomposition of the variance for overall satisfaction. First, we observe that over two thirds of total variance in SET is within-course. Given that the sampling 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 Fig. C1 in Appendix C we report, for each major, the ratio of the sampling to total variance at different sample sizes. The Figure shows that this hyperbolic relationship only plateaus at large values of course size. Therefore, while for vignette courses the problem is negligible, 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 far from negligible source of variability in SET.[16]

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

4.2. *Students' sorting across electives*

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 $(\alpha_i,\ \beta_i)$ 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 the available elective courses. Obviously, sets S_e are not disjoint, as students attend several electives, but their union coincides with the stratum. Since each vignette j is associated to one specific stratum, we count E_j electives that 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, $e = 1, \ldots, E_j$, we compare the average evaluation of vignette j provided by students in S_e, denoted $y_{j|e}$, with γ_j. As explained above, the latter is approximately equal to y_j. By using model (1), we have $y_{j|e} = \alpha_{\cdot|e} + \beta_{\cdot|e}\gamma_j + \varepsilon_{j|e}$, where $\alpha_{\cdot|e},\ \beta_{\cdot|e},\ \varepsilon_{j|e}$ are the averages of α_i, β_i and ε_{ij} conditional on S_e. If there is sorting, then $\alpha_{\cdot|e}$ and $\beta_{\cdot|e}$ will systematically differ from 0 to 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.[17]

Fig. 2 reports for each major (and pooling all strata within majors) the scatterplot of $y_{j|e}$ against the corresponding value of y_j. 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

[15] 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.

[16] Very similar results hold when the interest for the subject is removed from students' satisfaction (see Table C1 in Appendix C).

[17] In principle, one could also consider the option of correcting individual students' evaluations of elective course by means of the estimated $\widehat{\alpha}_i$ and $\widehat{\beta}_i$, 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.

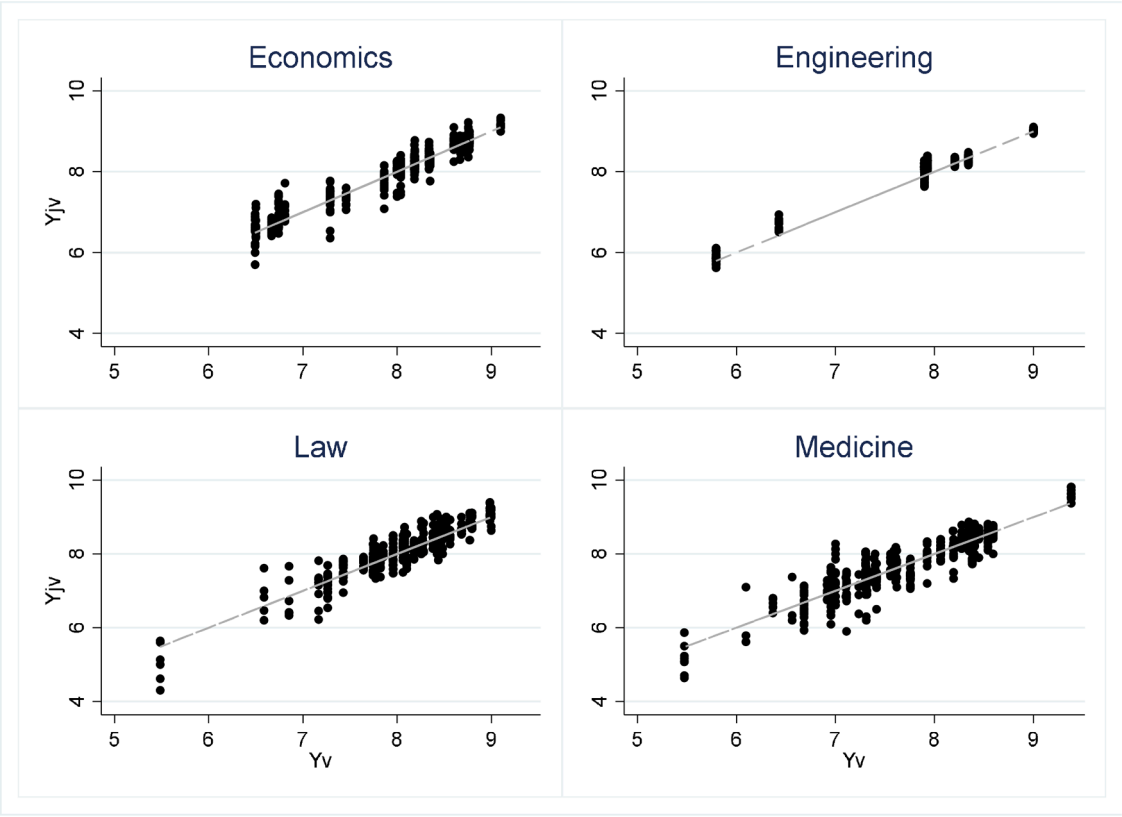

Fig. 2. Average evaluation of vignettes by students choosing elective *e*, $e = 1,\ldots, E$ vs overall average evaluation of vignettes – all strata pooled. By major.
Note: For each vignette *j*, the horizontal axis reports $y_{.j}$, the average evaluation provided by the students in the stratum which vignette *j* belongs to; the vertical axis reports $y_{.j|e}$, the average evaluation of vignette *j* provided by the students choosing elective *e*, $e = 1,\ldots,E_j$.

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.[18]

Was sorting absent, 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 separately for each major and test the null hypothesis of intercept equal to 0 and slope equal to 1.

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 rather imprecise and no firm conclusion can be established.[19] This evidence implies that the factors that make students prefer an elective over another are correlated with their reporting styles. This, in turn, makes it difficult to compare the SETs of different elective courses, because they embody different reporting styles.

We remark that the proposed test under-rejects the null of no-sorting because there are special instances where sorting on reporting style is compatible with zero intercept and unitary slope, i.e. instances in which this test has no power. 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 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.[20]

Table 4
Tests for sorting. By major.

	Economics (1)	Engineering (2)	Law (3)	Medicine (4)
Intercept	0.188**	0.340***	−0.561***	−0.149
	(0.095)	(0.095)	(0.187)	(0.168)
Slope	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*, $e = 1,\ldots,E_j$, 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.
*** $p < 0.01$,.
** $p < 0.05$, *$p < 0.1$.

Rejecting the null of no sorting does not necessarily imply that sorting on reporting styles is a concern of practical interest, as a large

[18] When the interest for the subject is removed from students' satisfaction results are comparable to those in Fig. 2 (See Fig. C2 in Appendix C).
[19] Results when satisfaction is net of the interest for the subject are in Table C2 in Appendix C and mirror those in Table 4.
[20] In our data, the union of S_e does not coincide 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.

Table 5
Individual characteristics and reporting heterogeneity. By major.

	Economics		Engineering		Law		Medicine	
	Coeff-	P-value	Coeff-	P-value	Coeff-	P-value	Coeff-	P-value
Constant	8.176	<0.001	7.439	<0.001	8.053	<0.001	7.663	<0.001
Female	−0.129	0.114	0.255	0.144	−0.040	0.653	−0.245	0.037
Local-born	−0.132	0.185	0.095	0.696	0.074	0.540	0.144	0.328
Year of birth	−0.099	0.805	0.126	0.912	−0.055	<0.001	0.066	0.248
High school grade	0.004	0.264	0.004	0.560	0.006	0.140	0.009	0.117
y_j	0.937	<0.001	1.180	<0.001	1.208	<0.001	1.093	<0.001
Female × y_j	0.305	<0.001	−0.048	0.745	0.146	0.358	0.421	0.002
Local − born × y_j	−0.179	0.081	−0.189	0.297	−0.319	0.139	−0.437	0.009
Year of birth × y_j	0.038	0.435	0.112	0.205	0.059	0.116	0.042	0.477
High school grade × y_j	−0.006	0.299	−0.001	0.919	−0.005	0.553	−0.001	0.822
Observations	1669		511		1662		1204	
Students	443		133		477		339	
Joint tests' p-values of heterogeneity in intercepts and slopes for:								
Female	0.003		0.337		0.648		0.005	
Local-born	0.020		0.557		0.329		0.032	
Year of birth	0.732		0.392		0.001		0.456	
High school grade	0.266		0.842		0.323		0.139	

Notes: the table reports the OLS estimates of Eq. (1), where we replace individual intercepts and slopes with four individual covariates (gender, birthplace, birthyear and high school grade) and their interactions with γ_j. To ease interpretation, year of birth, high school grade and γ_j are recentered to have zero mean. Inference is robust to clustering by student. Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

part of the deviations $y_{j|e} - y_j$ reported in Fig. 2 could still be attributable to sampling variability. To assess how relevant sorting is with respect to noise we define the ratio

$$S = \frac{\left(\mathrm{y}_j - \frac{1}{\mathrm{E}_j}\sum_e y_{j|e}\right)^2}{\frac{1}{\mathrm{E}_j}\sum_e \left(y_{j|e} - y_j\right)^2} \tag{6}$$

which is bounded between 0 and 1 and measures the intensity of sorting (see Appendix B).[21] Fig. C3 in Appendix C shows that the average $y_{j|e}$ does not coincide with y_j, and in several cases the deviation is substantial. Next, Fig. C4 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 vignettes for which sorting is predominant and S can even exceed 0.80. We conclude that, overall, sorting has practical relevance and cannot be neglected.

Finally, we investigate whether the available observables (gender, birthplace, birthyear, and high school grade) may account for reporting heterogeneity. For each observable, we partition students' population in two groups,[22] and we plot the average vignette evaluation by subgroup against the average evaluation in the population (see Figs. C5-C8 in Appendix C). Although the patterns vary slightly across majors, no variable appears to be a major driver of reporting styles.

To get firmer indications, we also estimate a modified version of Eq. (1), where we replace the individual intercepts and slopes with the four observables and their interactions with γ_j. Separately for each major, we then test the null hypothesis that the intercept and slope coefficients associated with each variable are jointly zero (see Table 5). Overall, results confirm that reporting heterogeneity cannot be easily attributed to any specific observable. In Economics and Medicine, female and local born students evaluate teaching differently than their male and non-local born counterparts, but this is not the case in Law and Engineering. In Engineering, younger students rate differently than older students, but not in the other three majors. Finally, in no major students with better high school grades rate their classes differently than students performing worse at high school. We conclude that richer data are needed to better explore the correlates of reporting heterogeneity.

4.3. Implications of reporting heterogeneity and noise

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.[23] Therefore, it is of policy relevance to gauge the extent to which such ranking is affected by 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 constructed as follows: for each stratum $t = 1,2,..,\mathrm{T_M}$ 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 because of reporting heterogeneity and the noise component – the latter being 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 Fig. 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 Fig. 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 $y_{j|e}$.[24]

[21] Ratio S is a lower-bound for the proportion of the dispersion of $y_{j|e}$ around y_j due to sorting. See Appendix B.

[22] Respectively, female vs male students, local born vs non-local born, birthyear and high school grade above and below median.

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

[24] Also in this case, results for satisfaction net of the interest for the subject – reported in Fig. C9 in Appendix C – are comparable to those in Fig. 3.

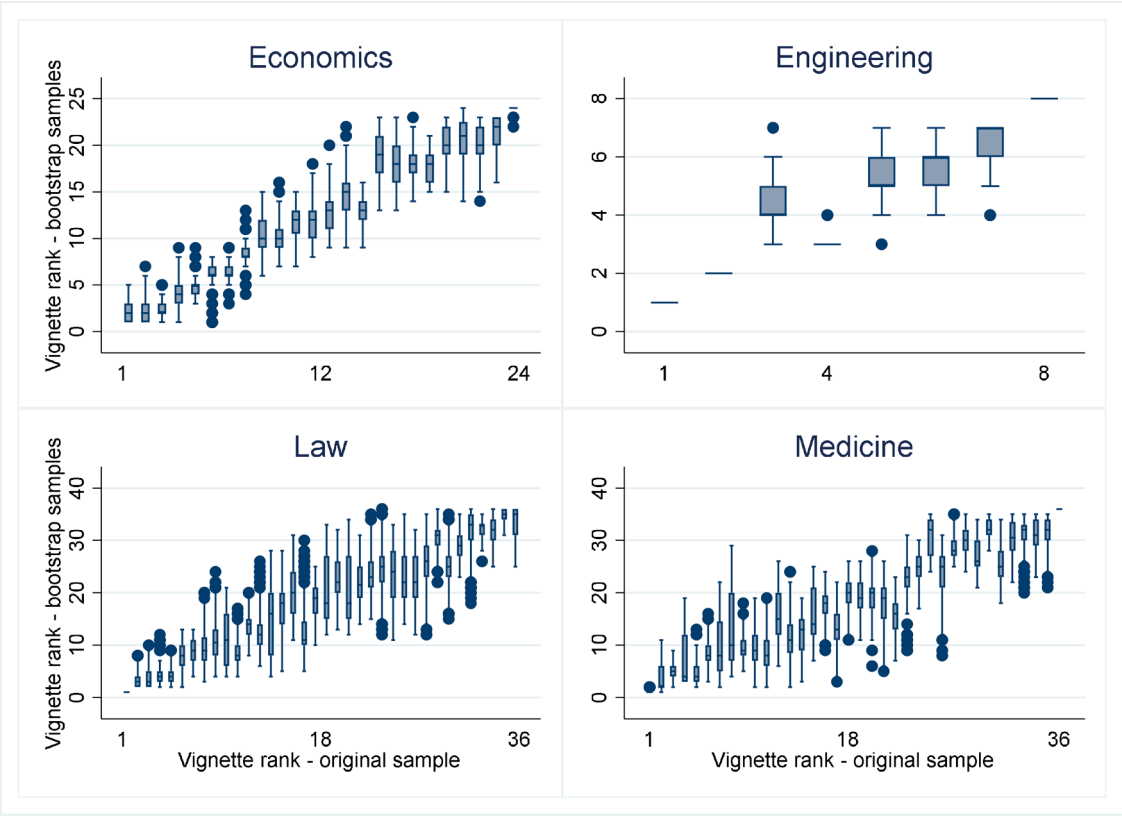

Fig. 3. Boxplots of bootstrapped rankings of courses. By major.
Note: 200 replications. In each replication, we randomly draw one elective course e per stratum, compute $y_{j|e}$ for all vignettes which belong to the stratum, pool all strata of the major and define the corresponding rank of each vignette. For each vignette the graph reports the boxplot of the distribution of the rank positions occupied by the vignette across the replications.

These results highlight how unreliable rankings based on SET can be, and how unfair it would be to construct incentive or base promotions on these rankings.

5. Correcting SET of elective courses

In principle, the availability of vignette courses offers the opportunity of anchoring students' evaluation of the electives and making them comparable. After all, this is precisely why vignettes have been proposed. In this section we provide a further note of caution because we show that, in practice, correcting evaluations might end up increasing rather than reducing the level of uncertainty about the true SET.

The argument proceeds as follows. Applying model (1) to elective course e, we get $y_{\cdot e} = \alpha_{\cdot|e} + \beta_{\cdot|e}\gamma_e + \varepsilon_{\cdot e}$, i.e. course e 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:

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

Since $\alpha_{\cdot|e}$ and $\beta_{\cdot|e}$ can be estimated by exploiting vignettes and regressing $y_{j|e}$ on γ_j, separately for each e, $y_{\cdot e}$ can be debiased. However, de-biasing $y_{\cdot e}$ comes at the price of inflating its sampling variability due to the fact that $\alpha_{\cdot|e}$ and $\beta_{\cdot|e}$ need to be estimated. Eventually, the increased uncertainty due to sampling variability might more than offset the benefit of removing the bias. To assess the trade-off between bias and variability, we compare the mean squared errors of $y_{\cdot e}$ and of its de-biased version $\widetilde{y}_{\cdot e}$.

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

$$MSE(y_{\cdot e}) = E(y_{\cdot e} - \gamma_e)^2 = \left[\alpha_{\cdot|e} + (\beta_{\cdot|e} - 1)\gamma_e\right]^2 + \frac{\sigma_\varepsilon^2}{n_e} \tag{8}$$

which is the sum of the squared bias and the sampling variance[25]

$$Var(y_{\cdot e}) = \frac{\sigma_\varepsilon^2}{n_e}. \tag{9}$$

Turning to $\widetilde{y}_{\cdot e}$, in Appendix B we show that the mean squared error $E(\widetilde{y}_{\cdot e} - \gamma_e)^2$ approximately coincides with the variance of $\widetilde{y}_{\cdot e}$, since $\widetilde{y}_{\cdot e}$ is approximately unbiased:

$$\begin{aligned} MSE(\widetilde{y}_{\cdot e}) &= Var(\widetilde{y}_{\cdot e}) \\ &= Var\left(\frac{y_{\cdot e} - \alpha_{\cdot|e}}{\beta_{\cdot|e}}\right) + Var\left(\frac{1}{\beta_{\cdot|e}}\left(\widehat{\alpha}_{\cdot|e} - \alpha_{\cdot|e}\right)\right) \\ &\quad + Var\left(\frac{y_{\cdot e} - \alpha_{\cdot|e}}{\beta_{\cdot|e}^2}\left(\widehat{\beta}_{\cdot|e} - \beta_{\cdot|e}\right)\right) \\ &\quad + 2cov\left(\frac{1}{\beta_{\cdot|e}}\left(\widehat{\alpha}_{\cdot|e} - \alpha_{\cdot|e}\right), \frac{y_{\cdot e} - \alpha_{\cdot|e}}{\beta_{\cdot|e}^2}\left(\widehat{\beta}_{\cdot|e} - \beta_{\cdot|e}\right)\right) \\ &= \frac{1}{\beta_{\cdot e}^2}\frac{\sigma_\varepsilon^2}{n_e}\left\{\frac{5}{4} + \frac{1}{\sum_{j=1}^4 \left(y_j - y_{\cdot\cdot}\right)^2}\left(\left(y_{\cdot\cdot} - \gamma_e\right)^2 + \frac{1}{\beta_{\cdot e}^2}\frac{\sigma_\varepsilon^2}{n_e}\right)\right\} \end{aligned} \tag{10}$$

Where $y_{\cdot\cdot}$ is the average of y_j over the four vignettes that can be linked to elective e.

Comparing (9) and (10), we note that the sampling variance of $\widetilde{y}_{\cdot e}$ is

[25] Assumption 1 implies that $\varepsilon_{\cdot e}$ is uncorrelated with γ_e, $\alpha_{\cdot|e}$ and $\beta_{\cdot|e}$.

certainly larger than that of $y_{\cdot e}$. It turns out that $MSE(\widetilde{y}_{\cdot e})$ 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 most cases, the additional sampling variance brought in by the de-biasing procedure exceeds the systematic bias of $y_{\cdot e}$.[26]

While this is clearly bad news regarding the usefulness of SET even when vignettes are available, we cannot exclude that the procedure suggested in this section could be applied more fruitfully in other contexts, where electives are larger and sampling variability is less of a concern.

6. 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 variability 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 de-bias SET and a criterion to decide whether to undertake it or not.

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, schools should be cautious in relying on teaching evaluations only when deciding on teachers' career. Stark and Freishtat (2014) argue that, at the bare 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 by which 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*, bearing in mind two crucial caveats: 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. 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.

CRediT authorship contribution statement

Marco Bertoni: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization. **Enrico Rettore:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization. **Lorenzo Rocco:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization.

Declarations of competing interest

None.

Data availability

We would be happy to share our data, but in case we must check if the university providing those data agree to share them.

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 Erich Battistin, Giorgio Brunello, David Card, Ingo Isphording, Ofer Malamud, Roberto Nisticò, Marco Paccagnella, Paolo Pinotti, Riccardo Saulle, Ulf Zoelitz, as well as from seminar participants at ANVUR in Rome, Bergamo, FBK-IRVAPP in Trento, Padova, Verona, ZEW in Mannheim, as well as at the EALE 2023 conference in Prague, and the IWAEE 2019 conference in Catanzaro. We acknowledge financial support from a CARIPARO foundation "Starting Grant". This paper uses confidential data from the archives of a large Italian University. We are willing to assist interested researchers in getting access to the data, and to share our STATA dofiles for replication. We have no relevant financial or material interest to disclose about this paper. All remaining errors are our own.

Appendix A: 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

[26] We also tried to improve the precision of the unbiased estimate $\widetilde{y}_{\cdot e}$ using empirical Bayes methods as in Kane and Staiger (2008), Chetty et al. (2014), Gilraine et al. (2020), among others. Unfortunately, for most of the electives included in our study the resulting estimate is only marginally more precise than the corresponding $\widetilde{y}_{\cdot e}$. This happens because, for most electives, the variance of the *a priori* distribution of γ_e is much larger than the sampling variance of $\widetilde{y}_{\cdot e}$. As a result, the shrinkage estimator is dominated by $\widetilde{y}_{\cdot e}$.

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.[27]

As reported in Table A1, we retain 598 students in Economics; 242 in Engineering; 1317 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 the second row of Table A1, 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. Since we can only rely on vignette responses – not affected by sorting – to estimate student response styles, we further retain only students who evaluate at least three out of the four vignettes defined for their stratum.[28] This requirement implies a substantial reduction in the sample of students, which is necessarily more severe in the degrees of Law and Medicine. 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 Fig. 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.

Table A1
Derivation of the study sample.

	Economics			Engineering			Law			Medicine		
	Students (1a)	Courses (1b)	Strata (1c)	Students (2a)	Courses (2b)	Strata (2c)	Students (3a)	Courses (3b)	Strata (3c)	Students (4a)	Courses (4b)	Strata (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 (1a)	Final sample (1b)	Reference population (2a)	Final sample (2b)	Reference population (3a)	Final sample (3b)	Reference population (4a)	Final sample (4b)	Reference population (5a)	Final sample (5b)

(*continued on next page*)

[27] 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.

[28] 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.

Table A2 (*continued*)

	Number of students		Female		Local-born student		Year of birth (19-)		High school grade (60–100)	
	Reference population (1a)	Final sample (1b)	Reference population (2a)	Final sample (2b)	Reference population (3a)	Final sample (3b)	Reference population (4a)	Final sample (4b)	Reference population (5a)	Final sample (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

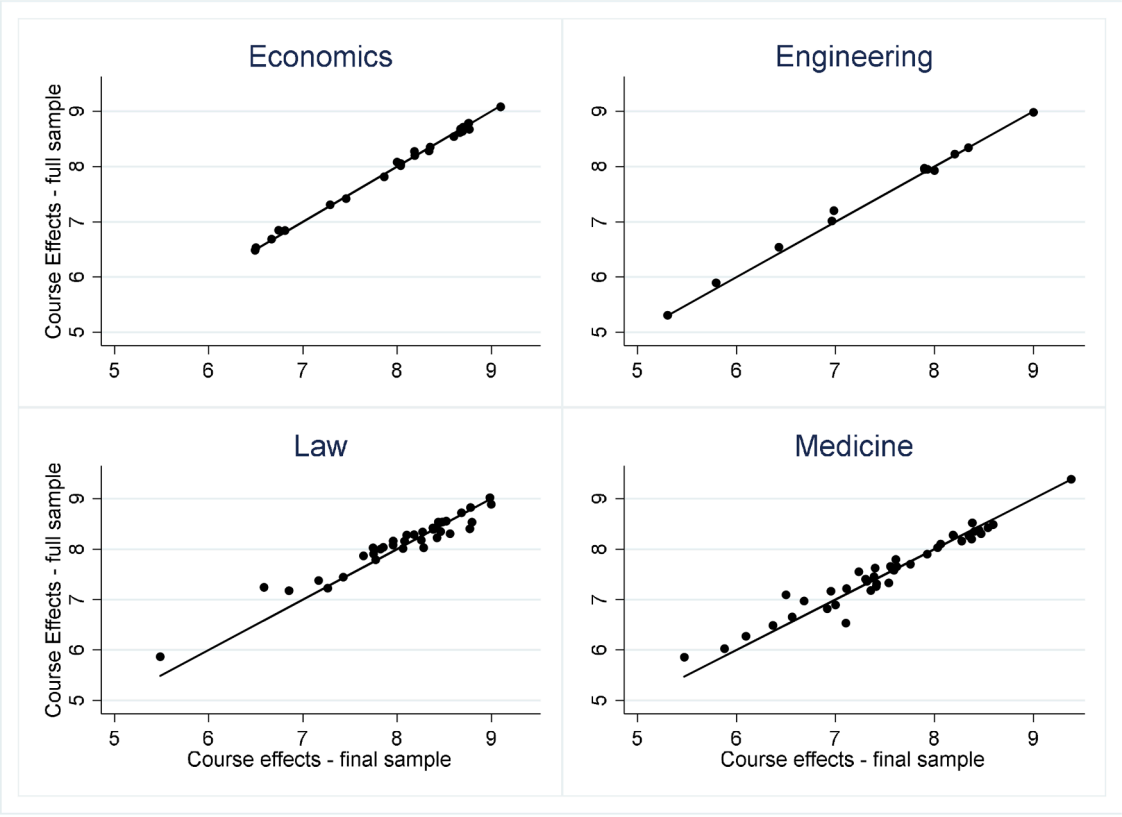

Fig. A1. Average evaluation of vignettes in the reference population vs final sample – including also dropped strata. By major.

Appendix B. Derivation of Eqs. (6) and (10)

Eq. (6)

The deviation $y_{j|e} - y_{.j}$ can be written as $y_{j|e} - y_{.j} = \alpha_{.|e} + \left(\beta_{.|e} - 1\right)\gamma_j + \varepsilon_{j|e}$. For each vignette j, the average of the squared deviation[29] is

$$\frac{1}{\mathrm{E}_j}\sum_e \left(y_{j|e} - y_{.j}\right)^2 = \frac{1}{\mathrm{E}_j}\sum_e \varepsilon_{j|e}^2 + \frac{1}{\mathrm{E}_j}\sum_e \left(\alpha_{.|e} + (\beta_{.|e} - 1)\gamma_j\right)^2 =$$
$$= \frac{1}{\mathrm{E}_j}\sum_e \varepsilon_{j|e}^2 + Var\left(\alpha_{.|e} + (\beta_{.|e} - 1)\gamma_j\right) + \left[E\left(\alpha_{.|e} + (\beta_{.|e} - 1)\gamma_j\right)\right]^2 \quad \text{(B1)}$$

If the allocation of students across electives was random, then $y_{j|e} \simeq y_{.j} \simeq \gamma_j$, and the term $\left[E\left(\alpha_{.|e} + (\beta_{.|e} - 1)\gamma_j\right)\right]^2$ would vanish.[30] The latter

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

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 $\left(y_{.j} - \frac{1}{E_j}\sum_e y_{.j|e}\right)^2$.

Hence, the ratio $S = \frac{\left(y_{.j} - \frac{1}{E_j}\sum_e y_{.j|e}\right)^2}{\frac{1}{E_j}\sum_e \left(y_{.j|e} - y_{.j}\right)^2}$ is an index, defined between 0 and 1, which measures sorting intensity.[31]

Eq. (10)

Let $\widehat{\alpha}_{\cdot|e}$ and $\widehat{\beta}_{\cdot|e}$ be the estimates from the regression $y_{\cdot e} = \alpha_{\cdot|e} + \beta_{\cdot|e}\gamma_e + \varepsilon_{\cdot e}$. Equipped with such estimates, we can compute $\widetilde{y}_{\cdot e} = \frac{y_{\cdot e} - \widehat{\alpha}_{\cdot|e}}{\widehat{\beta}_{\cdot|e}}$. A first order Taylor expansion of $\widetilde{y}_{\cdot e}$ around $\alpha_{\cdot|e}$ and $\beta_{\cdot|e}$ yields:

$$\widetilde{y}_{\cdot e} \sim \frac{y_{\cdot e} - \alpha_{\cdot|e}}{\beta_{\cdot|e}} - \frac{1}{\beta_{\cdot|e}}\left(\widehat{\alpha}_{\cdot|e} - \alpha_{\cdot|e}\right) - \frac{y_{\cdot e} - \alpha_{\cdot|e}}{\beta^2_{\cdot|e}}\left(\widehat{\beta}_{\cdot|e} - \beta_{\cdot|e}\right) \tag{B2}$$

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

Then, $MSE(\widetilde{y}_{\cdot e})$ is approximately equal to:

$$MSE(\widetilde{y}_{\cdot e}) = Var(\widetilde{y}_{\cdot e}) = Var\left(\frac{y_{\cdot e} - \alpha_{\cdot|e}}{\beta_{\cdot|e}}\right) + Var\left(\frac{1}{\beta_{\cdot|e}}\left(\widehat{\alpha}_{\cdot|e} - \alpha_{\cdot|e}\right)\right) + Var\left(\frac{y_{\cdot e} - \alpha_{\cdot|e}}{\beta^2_{\cdot|e}}\left(\widehat{\beta}_{\cdot|e} - \beta_{\cdot|e}\right)\right) + 2cov\left(\frac{1}{\beta_{\cdot|e}}\left(\widehat{\alpha}_{\cdot|e} - \alpha_{\cdot|e}\right), \frac{y_{\cdot e} - \alpha_{\cdot|e}}{\beta^2_{\cdot|e}}\left(\widehat{\beta}_{\cdot|e} - \beta_{\cdot|e}\right)\right) \tag{B3}$$

This holds because $cov\left(\frac{y_{\cdot e} - \alpha_{\cdot|e}}{\beta_{\cdot|e}}, \frac{1}{\beta_{\cdot|e}}\left(\widehat{\alpha}_{\cdot|e} - \alpha_{\cdot|e}\right)\right) = 0$ and $cov\left(\frac{y_{\cdot e} - \alpha_{\cdot|e}}{\beta_{\cdot|e}}, \frac{y_{\cdot e} - \alpha_{\cdot|e}}{\beta^2_{\cdot|e}}\left(\widehat{\beta}_{\cdot|e} - \beta_{\cdot|e}\right)\right) = 0$. Intuitively, for given e, the sampling error embodied in the estimates $\widehat{\alpha}_{\cdot|e}$ and $\widehat{\beta}_{\cdot|e}$ 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:

$$MSE(\widetilde{y}_{\cdot e}) = \frac{1}{\beta^2_{\cdot e}}\frac{\sigma^2_\varepsilon}{n_e}\left\{\frac{5}{4} + \frac{1}{\sum_{j=1}^4 \left(y_{\cdot j} - y_{\cdot\cdot}\right)^2}\left(\left(y_{\cdot\cdot} - \gamma_e\right)^2 + \frac{1}{\beta^2_{\cdot e}}\frac{\sigma^2_\varepsilon}{n_e}\right)\right\} \tag{B4}$$

where $y_{\cdot\cdot}$ is the average of $y_{\cdot j}$ over the four vignettes.

Appendix C. Additional Tables and Figures

Table C1
Decomposition of the variance of SET purged of interest for the subject 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	0.292	0.708	0.709	0.321
Engineering	0.288	0.712	0.571	0.429
Law	0.163	0.837	0.777	0.223
Medicine	0.195	0.805	0.733	0.267

Table C2
Tests for sorting. SET purged of interest for the subject. By major.

	Economics (1)	Engineering (2)	Law (3)	Medicine (4)
Intercept	0.031***	0.084***	0.014	−0.019
	(0.010)	(0.013)	(0.015)	(0.016)
Slope	0.976***	0.957***	1.068***	1.025***
	(0.012)	(0.012)	(0.025)	(0.022)
Observations	492	144	376	436

(*continued on next page*)

[31] Ratio S is a lower-bound for the proportion of the dispersion of $y_{j|e}$ around y_j due to sorting. With sorting, also the component $Var\left(\alpha_{\cdot|e} + \left(\beta_{\cdot|e} - 1\right)\gamma_j\right)$ of the second line of equation (B1) increases.

Table C2 (*continued*)

	Economics (1)	Engineering (2)	Law (3)	Medicine (4)
R-squared	0.933	0.977	0.833	0.833
P-values for:				
H0: $\alpha = 0$	0.002	<0.001	0.328	0.228
H0: $\beta = 1$	0.040	<0.001	0.006	0.260
H0: $(\alpha = 0; \beta = 1)$	0.006	<0.001	0.002	0.382

Note: see Table 4.

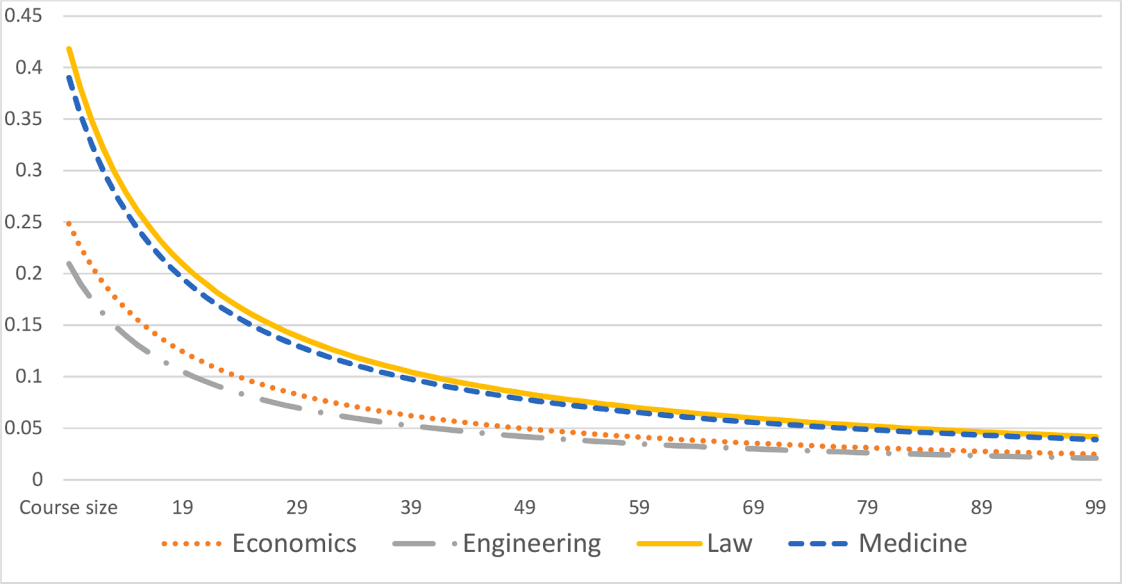

Fig. C1. Sampling over total variance as course size increases (see main text). By major.

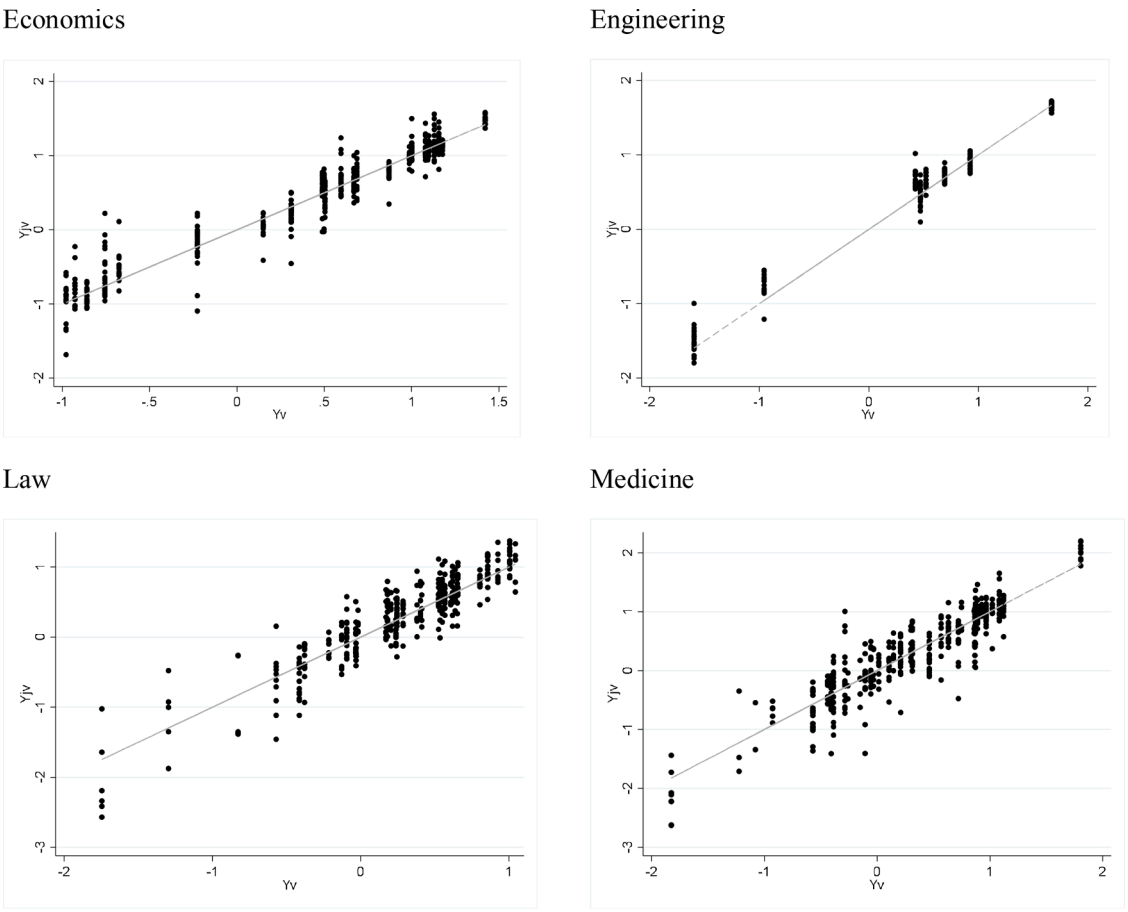

Fig. C2. Average evaluation of vignettes by students choosing elective e, $e = 1,\ldots, E$ vs overall average evaluation of vignettes – all strata pooled. SET purged of interest for the subject. By major.
Note: see Fig. 2.

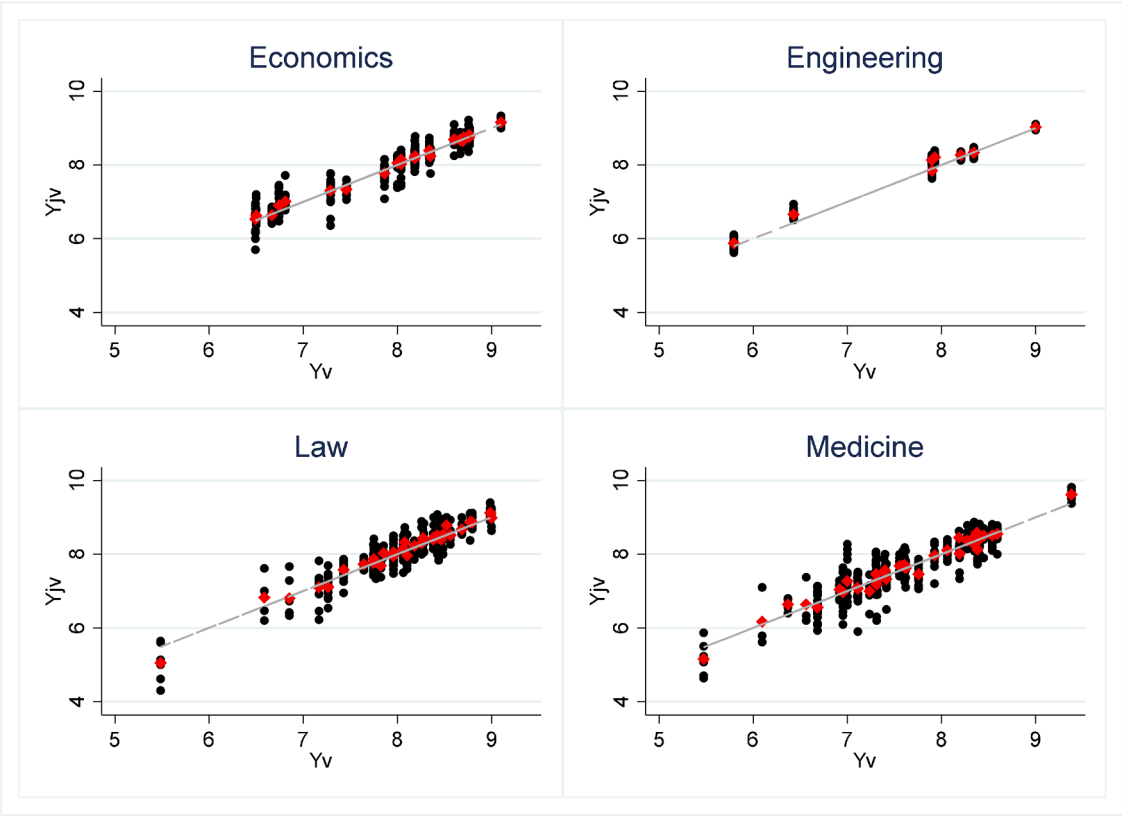

Fig. C3. Dispersion of y_{jle} and the average of y_{jle} for all vignettes. By major.
Note: red dots correspond to the average of y_{jle}, by vignette. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

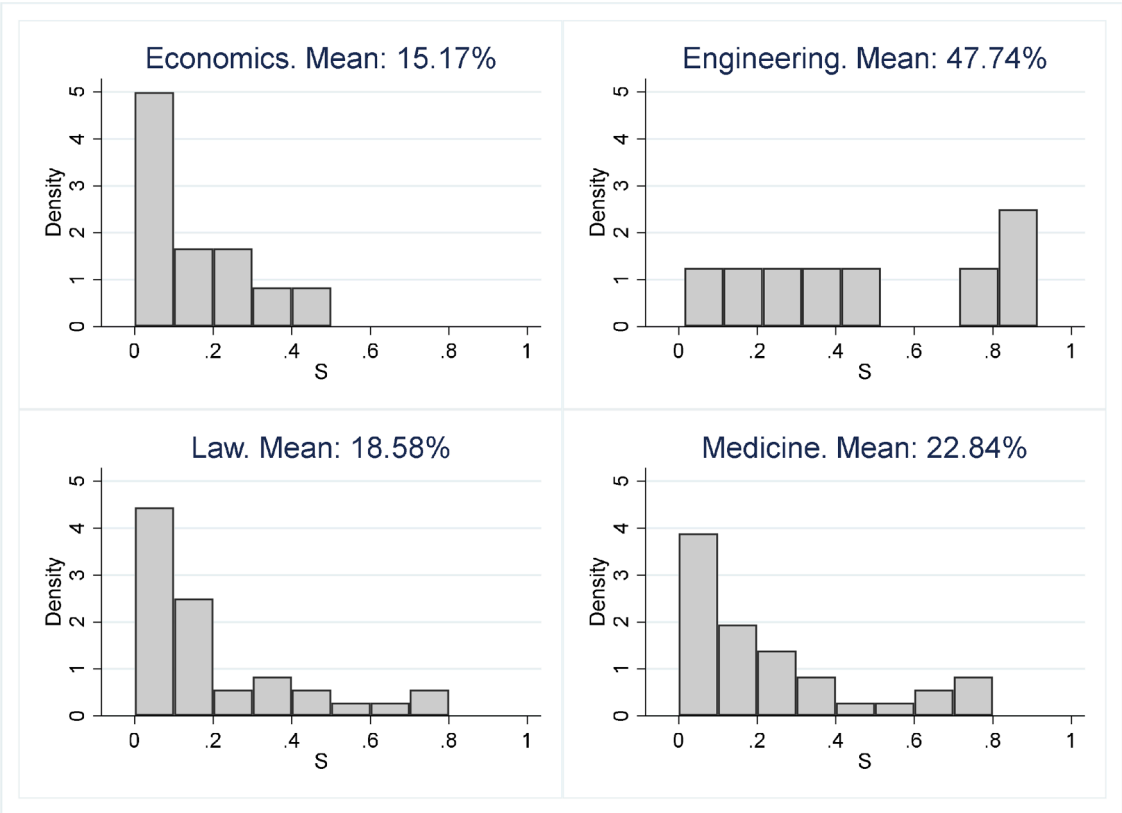

Fig. C4. Distribution of ratio S, the importance of sorting in vignette evaluation, across vignettes. By major.
Note: histogram bins have width 0.1.

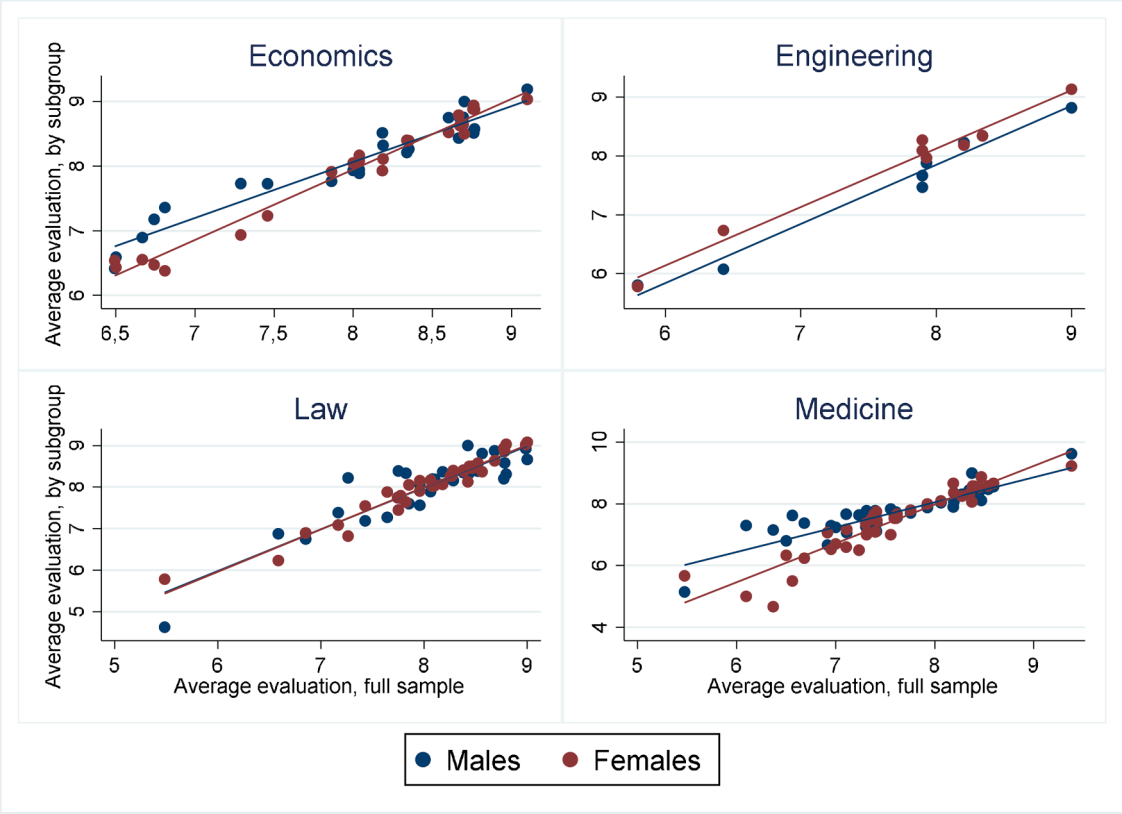

Fig. C5. Average vignette evaluation by group: male and female students.

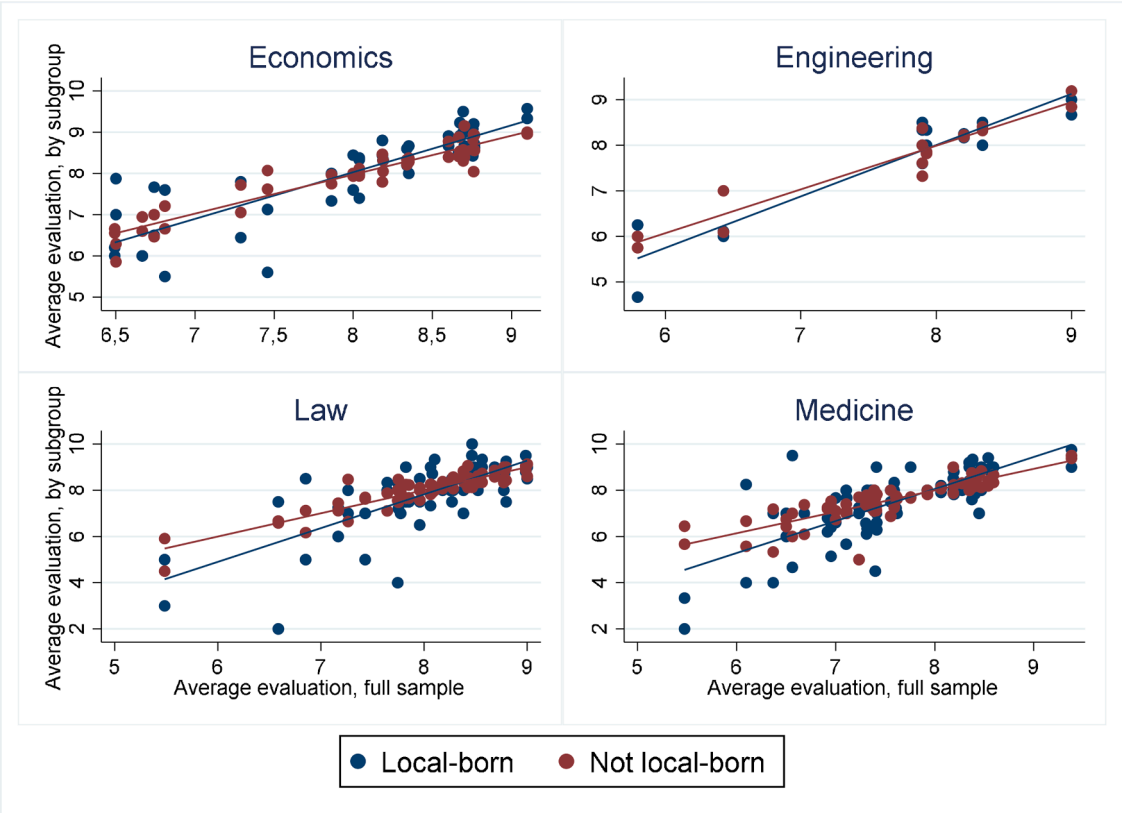

Fig. C6. Average vignette evaluation by group: local and non-local born students.

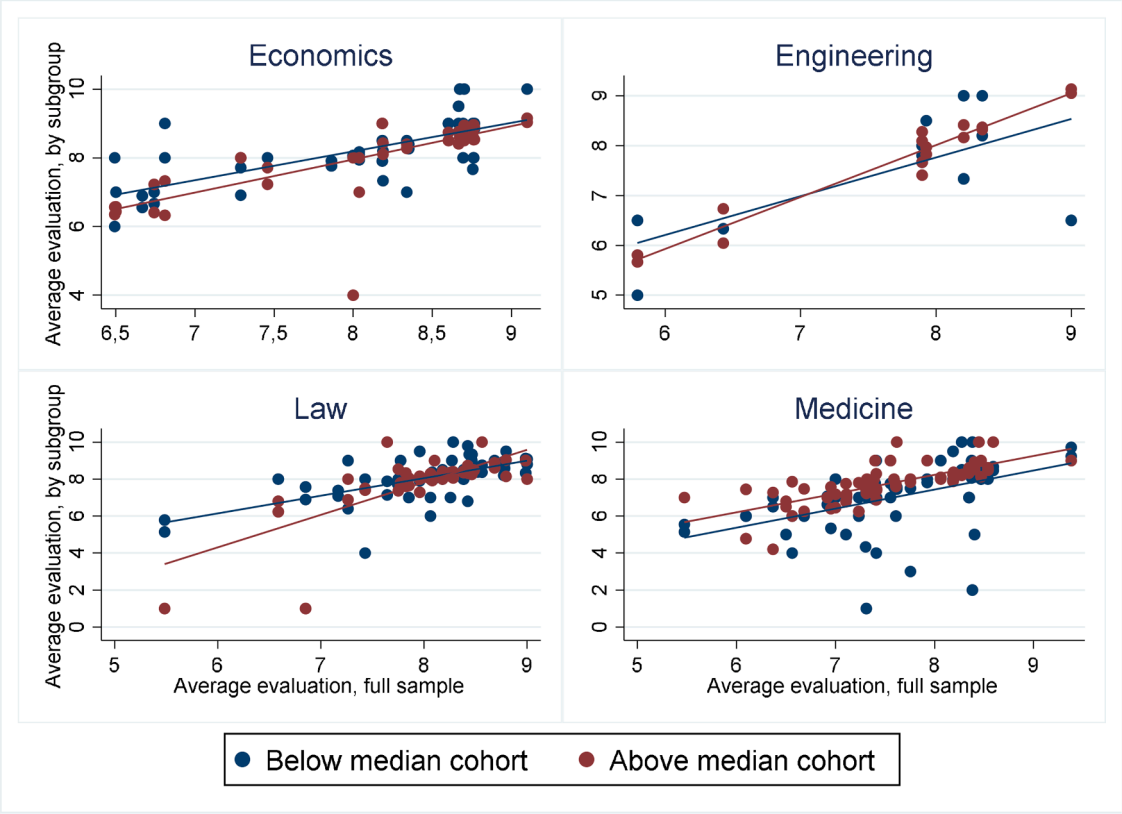

Fig. C7. Average vignette evaluation by group: below and above median birthyear.

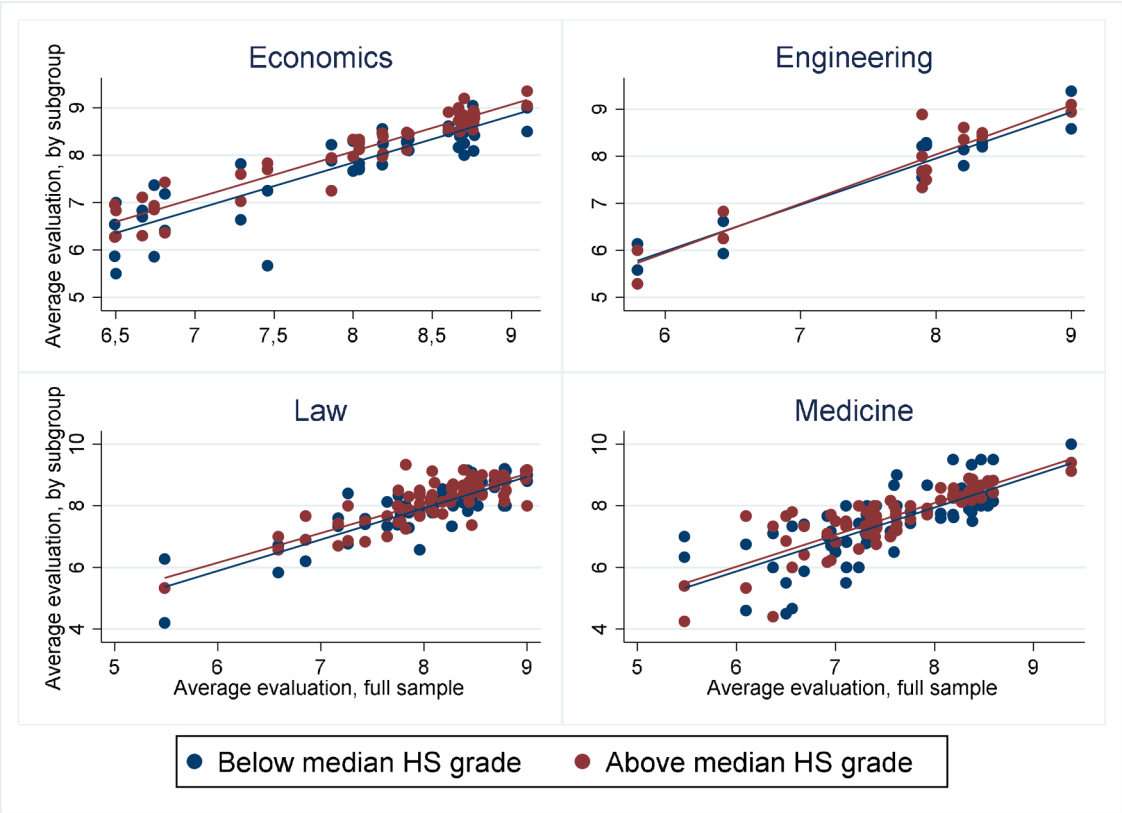

Fig. C8. Average vignette evaluation by group: below and above median high school grade.

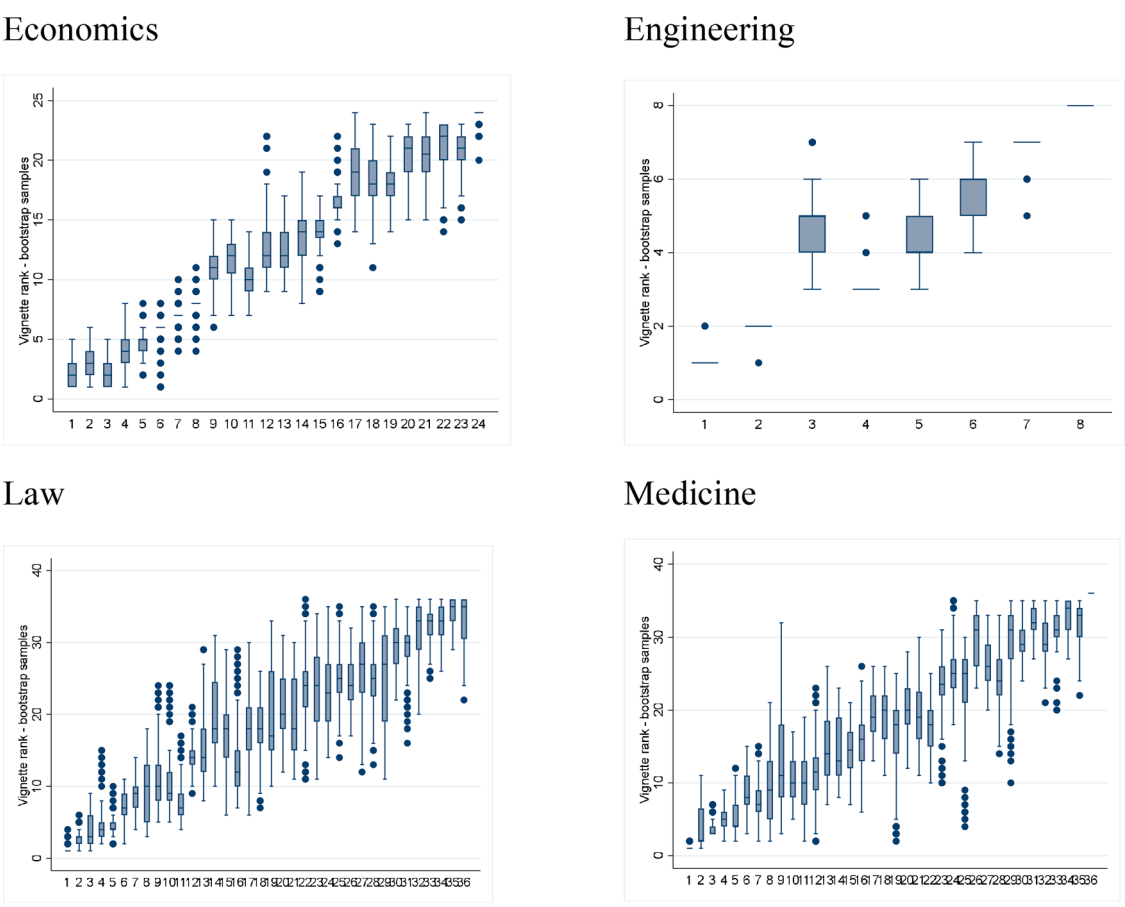

Fig. C9. Boxplots of bootstrapped rankings of courses. SET purged of interest for the subject. By major.
Note: see Fig. 3.

References

Becker, W.E., Watts, M., 1999. How departments of economics evaluate teaching. Am. Econ. Rev. 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. Rev. Econ. Statistics 95 (5), 1468–1479.

Bond, T.N., Lang, K., 2019. The sad truth about happiness scales. J. Politic. Econ. 127 (4), 1629–1640.

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

Boring, A., 2017. Gender biases in student evaluations of teachers. J. Public Econ. 145, 27–41.

Braga, M., Paccagnella, M., Pellizzari, M., 2014. Evaluating students' evaluations of professors. Econ. Educ. Rev. 41, 71–88.

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

Chetty, R., Friedman, J.N., Rockoff, J.E., 2014. Measuring the impacts of teachers I: evaluating bias in teacher value-added estimates. Am. Econ. Rev. 104 (9), 2593–2632.

Gilraine, M., Gu, J., and McMillan, R., 2020. A new method for estimating teacher value-added. NBER working paper 27084.

Goos, M., Salomons, A., 2017. Measuring teaching quality in higher education: assessing selection bias in course evaluations. Res. High. Educ. 58, 341–364.

Hamermesh, D.S., Parker, A., 2005. Beauty in the classroom: instructors' pulchritude and putative pedagogical productivity. Econ. Educ. Rev. 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. Med. Educ. 52 (10), 1064–1072.

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

Huettner, F., Sunder, M., 2012. Axiomatic arguments for decomposing goodness of fit according to Shapley and Owen values. Electron. J. Stat. 6, 1239–1250.

Kane, T.J., Staiger, D.O., 2008. Estimating teacher impacts on student achievement: an experimental evaluation. NBER Working Paper 14607.

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.

de Koning, B.K., Kunn-Nelen, A., Kunn, S., 2022. Student Satisfaction Scores Affect Enrollment in Higher Education Programs. WP Maastricht University.

Linask, M., Monks, J., 2018. Measuring faculty teaching effectiveness using conditional fixed effects. J. Econ. Educ. 49 (4), 324–339.

Langbein, L., 2008. Management by results: student evaluation of faculty teaching and the mis-measurement of performance. Econ. Educ. Rev. 27 (4), 417–428.

MacNell, L., Driscoll, A., Hunt, A.N., 2015. What's in a name: exposing gender bias in student ratings of teaching. Innov. High. Educ. 40 (4), 291–303.

McPherson, M.A., 2006. Determinants of how students evaluate teachers. J. Econ. Educ. 37 (1), 3–20.

Mengel, F., Sauermann, J., Zölitz, U., 2019. Gender bias in teaching evaluations. J. Eur. Econ. Assoc. 17 (2), 535–566.

Ponzo, M., Scoppa, V., 2013. Professors' Beauty, Ability, and Teaching Evaluations in Italy. B e J. Econom. Anal. Policy. 13 (2), 811–835.

Rivera, L.A., Tilcsik, A., 2019. Scaling down inequality: rating scales, gender bias, and the architecture of evaluation. Am. Sociol. Rev. 84 (2), 248–274.

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

Spooren, P., Brockx, B., Mortelmans, D., 2013. On the validity of student evaluation of teaching: the state of the art. Rev. Educ. Res. 83 (4), 598–642.

Stark, P.B., Freishtat, R., 2014. An evaluation of course evaluations. ScienceOpen. Res. https://doi.org/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. Econ. Educ. Rev. 54 (54), 7994.

Weinberg, B.A., Hashimoto, M., Fleisher, B.M., 2009. Evaluating teaching in higher education. J. Econ. Educ. 40 (3), 227–261.

Wolbring, T., Treischl, E., 2015. Selection bias in students' evaluation of teaching. Res. High. Educ. 1–21.