Reporting Heterogeneity and Noise in Student Evaluations of Teaching

Marco Bertoni (UNIPD)

Enrico Rettore (UNITN & FBK-IRVAPP)

Lorenzo Rocco (UNIPD)

April 2019

What are Students' Evaluation of Teaching (SET)?

Overall, how satisfied are you with this course?

1 2 3 4 5 6 7 8 9 10

Motivation

- Student Evaluations of Teaching (SET) were introduced in the 1920’s to provide feedback to instructors about their teaching practices

- Now performed in many universities around the world, with a broader purpose, often listed among the elements used to decide promotions in academia

- In Italy, loose reference to SET also in the *Decreto Legislativo 19/2012 - “Accreditamento, Valutazione periodica, Autovalutazione”* . Not unrealistic that SET will enter the set of relevant information for public choices in the field of higher education

Troubles with SET validity – the literature

- Several negative findings
    - Myopic students reward instructors awarding higher grades in the short run (Carrell and West, 2010, Braga *et al*, 2014)
    - SET are affected by the physical appearance of the instructors (Hamermesh and Parker, 2005, Ponzo and Scoppa, 2013)
    - There is gender discrimination (Boring, 2017) – even in questions unrelated to teaching quality (Mengel *et al*, 2018)
    - Non-response bias is a serious issue (Goose and Salmon, 2017, Spooren and Van Loon, 2012)

Troubles with SET validity – our take

1. *Noise:* low precision in the estimation of average SET by course

2. *Heterogeneity in response styles* may hamper comparability of SET if students with different response styles *sort* into different courses

   - Well-known problem in social sciences – plagues comparability of subjective measures of happiness, political efficacy, health, …
   - Never addressed so far for the case of SET

Noise

- Standard assumption in the use of SET: students are «unbiased» evaluators.

$$y_{ij} = \gamma_j + \varepsilon_{ij}$$

ε_{ij} is classical measurement error, with $\mathrm{E}(\varepsilon_{ij}) = 0$ and uncorrelated with γ_j

- The average course evaluation is $y_{.j} = \gamma_j + \varepsilon_{.j}$
- If the variance of measurement error is high with respect to the total variance of y_{ij}, point estimates of average course quality can be estimated with large standard errors

Heterogeneity in response styles

- We can characterize students' response styles in terms of two features:

  1. Differences in «*level*» – how lenient/strict an evaluator is on average
  2. Differences in «*slope*» – how sensitive an evaluator is to differences in quality

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

- Unbiased evaluators have $y_{ij} = 0 + 1 \times \gamma_j + \varepsilon_{ij}$
- *Level* effect: generous evaluators have $\alpha_i > 0$, severe ones have $\alpha_i < 0$
- *Slope* effect: hyper-sensitive evaluators have $\beta_i > 1$, hypo-sensitive ones have $\beta_i < 1$

Sorting

- Heterogeneity in response styles would not be a problem if each course was evaluated by (all or) a random sample of students
- On average, their evaluations would be unbiased
- But students self-sort into elective courses
- Similarly, sorting would not be a problem in absence of heterogeneity

A possible solution – anchoring vignettes (King, 2004)

How satisfied are you with your life in general?

Very Satisfied	Satisfied	Neither satisfied Nor dissatisfied	Dissatisfied	Very Dissatisfied
☐1	☐2	☐3	☐4	☐5

John is 63 years old. His wife died 2 years ago and he still spends a lot of time thinking about her. He has 4 children and 10 grandchildren who visit him regularly. John can make ends meet but has no money for extras such as expensive gifts to his grandchildren. He has had to stop working recently due to heart problems. He gets tired easily. Otherwise, he has no serious health conditions.

How satisfied with his life do you think John is?

Source: SHARE - wave 2

Our approach

- We *do not* have anchoring vignettes in our survey
- Compulsory courses, evaluated by all students within majors/cohorts/tracks, play the role of vignettes
- No sorting + large number of evaluations → $y_{.j} = \gamma_j$

- We use administrative data that allows to link *all* the scores assigned by a specific student to the courses she attended

- Each student evaluates multiple vignettes, providing us with enough within-student variation to disentangle reporting heterogeneity, noise and genuine differences between-courses

Our findings

- At most one third of individual SET variance is between courses. Of the within-course variance, 25% to 45% is due to *heterogeneity*, and the remaining part to *noise.*

- There is significant evidence that students with different reporting styles sort across elective courses

- Using a simulation exercise, we show dramatic consequences of noise/sorting for rakings of courses *within major*.

Implications

- Reporting heterogeneity and sorting – on top of noise - hamper the comparability of the average evaluations of courses attended by different subsets of students within majors.

- SET should not be used to incentivise, promote or hire teachers, especially within tournament-like schemes.

Our Data

- Administrative SET archive for four degree courses in a large Italian University:
  - *Laurea (3 years)* in Economics
  - *Laurea a ciclo unico (5 years)* in Architecture & Construction Engineering
  - *Laurea a ciclo unico (5 years)* in Law
  - *Laurea a ciclo unico (6 years)* in Medicine
- 3 years of SETs: 2011/12 to 2013/14 and 3 cohorts of students: matriculation in October 2011 to 2013

- We can link all evaluations provided by a given student

- Students within a course and cohort may be divided in tracks
- Define as «stratum» the combination of major, cohort and track
- We define each «course» as a learning unit taught by a specific professor to students belonging to a given stratum

Vignette courses

- We need to define for each stratum a set of compulsory courses evaluated by (close to) all students, that will serve as anchoring vignettes
- We choose the four courses with the highest coverage by stratum

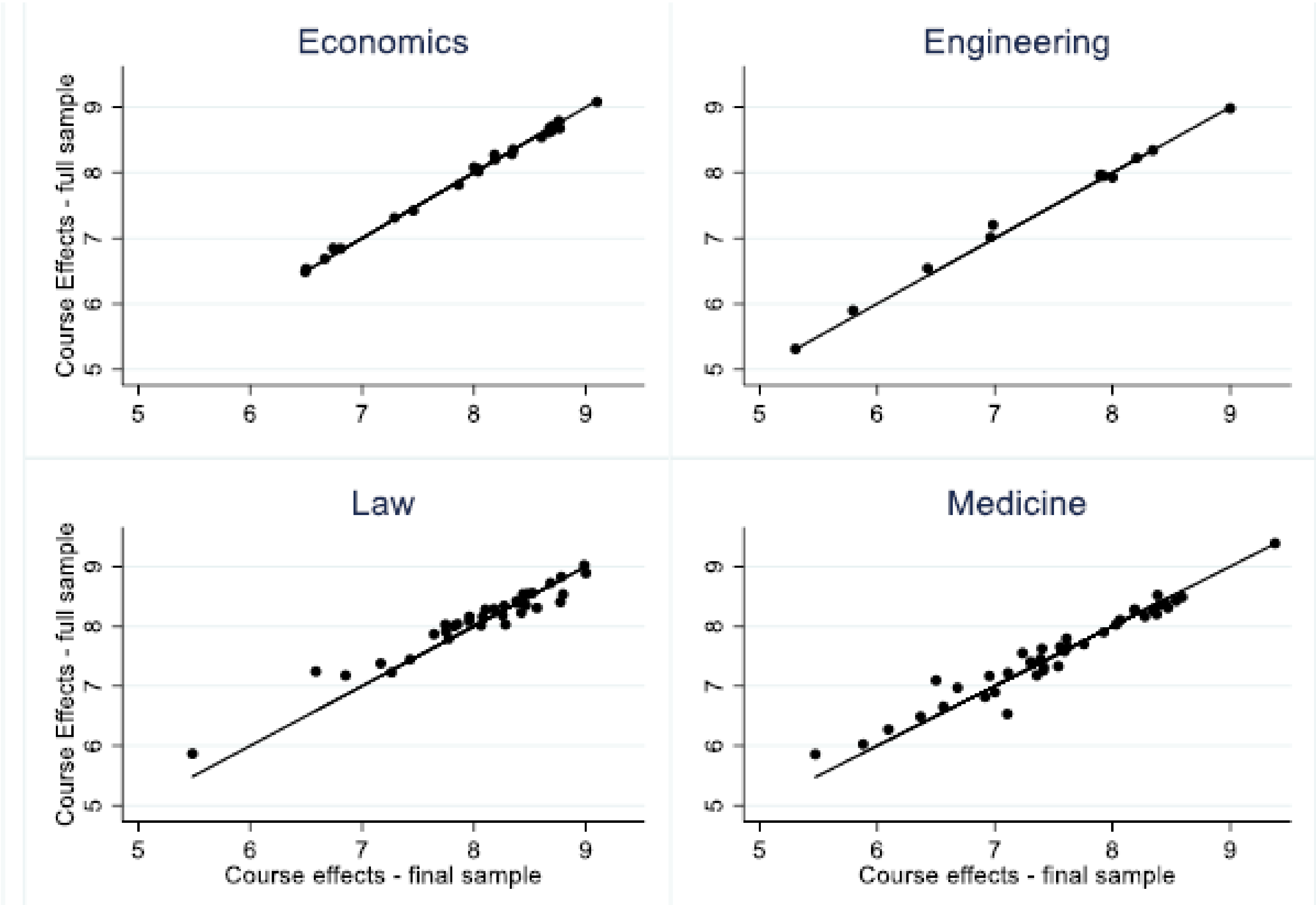
- Our analysis is demanding in terms of:
  - Number of vignette evaluations per student
  - Variation in vignette evalutions within students
  - Variation in average vignette evaluations within strata
  - Size of elective courses that we consider

- As a result, complex and demanding sample selection criteria
- Strict tests for representativeness – we drop strata that do not pass them

Sample selection: criteria and consequences

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

Sample selection: vignette evaluations



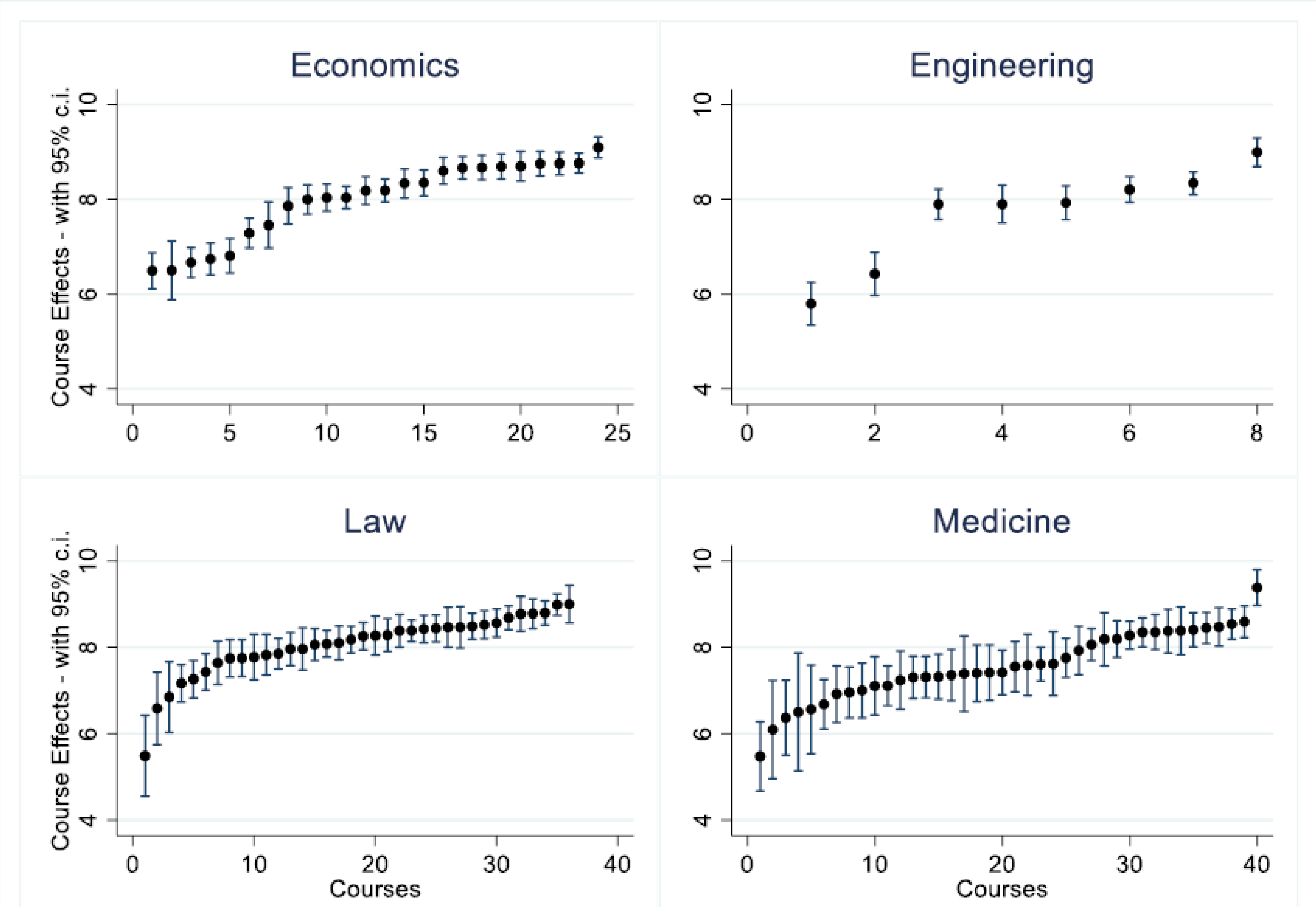
Sample selection: observables

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

The study sample: students and courses

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

Vignette course effects (with 95% c.i.)



What are the drivers of overall satisfaction?

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

Variance decomposition (1)

- Decomposing TOTAL variance of SET with reference to vignettes:

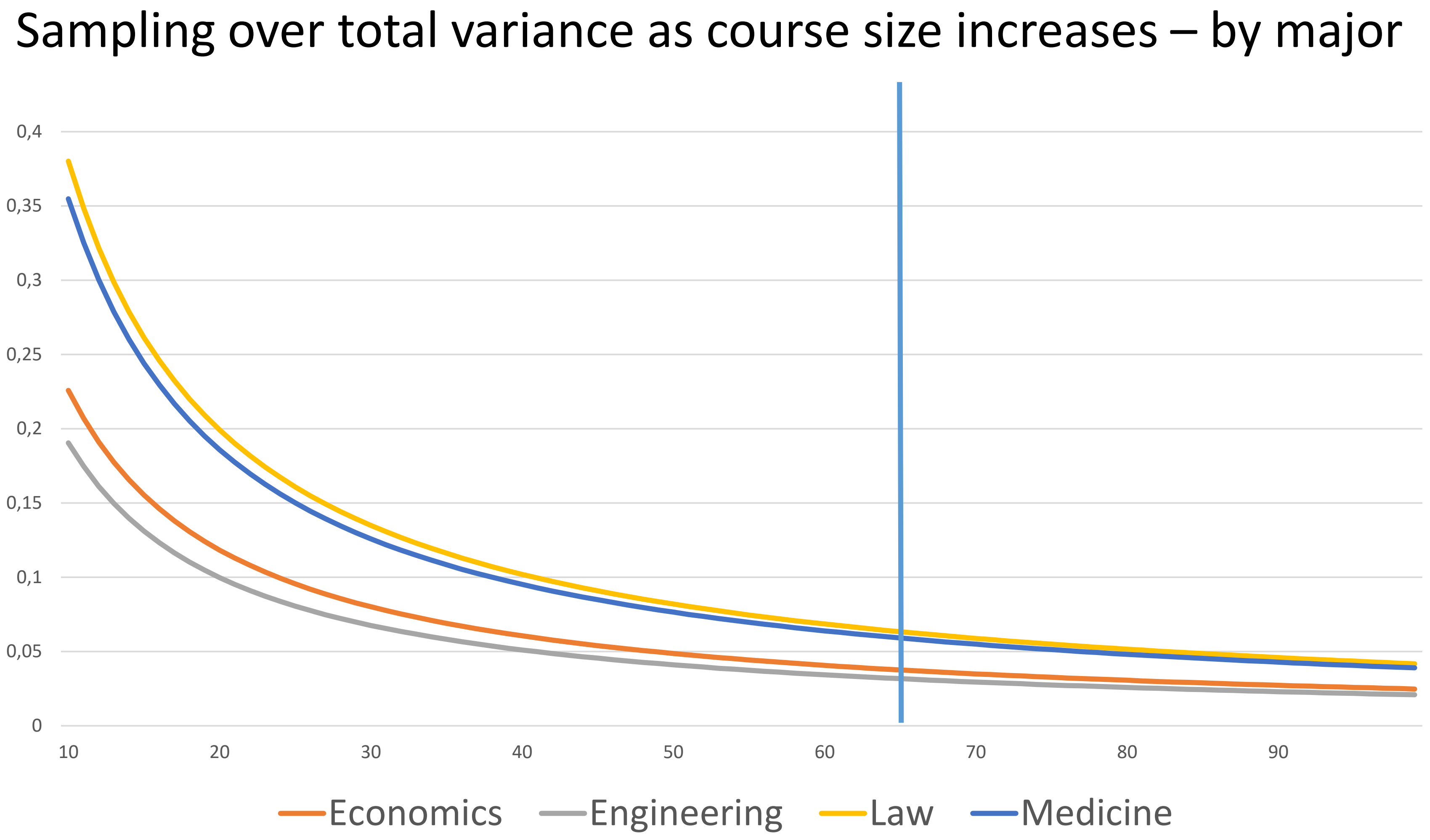
VARIANCE BETWEEN COURSES + VARIANCE WITHIN COURSES

- Variance between courses reflects genuine differences in course quality
- If heterogeneity in response styles were absent, all variability within courses would be due to noise

Variance decomposition (1)

	Variance between courses	Variance within courses
	% of total variance	% of total variance
	(1)	(2a)
Economics	0.287	0.713
Engineering	0.323	0.677
Law	0.193	0.807
Medicine	0.204	0.796

Variance decomposition (1) - implications



Variance decomposition (2)

- Students evaluating different electives can also differ in their response styles

- We estimate $y_{ij} = \alpha_i + \beta_i \gamma_j + \varepsilon_{ij}$ on individual-level vignette evaluations
  - Simple OLS model, exploiting that for vignettes $y_{.j} = \gamma_j$
  - α_i and β_i are estimated out of 3-4 observations per student, very noisy: carefully trim outliers

- By so doing, we get an estimate of σ_ε^2
- We can further decompose the variance WITHIN courses in two components:

REPORTING HETEOGENEITY (α_i and β_i) VS. NOISE (ε_{ij})

Variance decomposition (2)

	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

Testing for sorting on reporting styles: the problem

- Given the relevance of reporting heterogeneity, it becomes important to assess whether students sort across elective courses depending on their reporting styles

- In principle, to assess the bias induced by sorting we should compare

the **observed** average evaluation of a specific elective for the students who **actually evaluated** it	VS.	the **counterfactual** average evaluation of that same elective if **all** students evaluated it

- **This is something we cannot do**

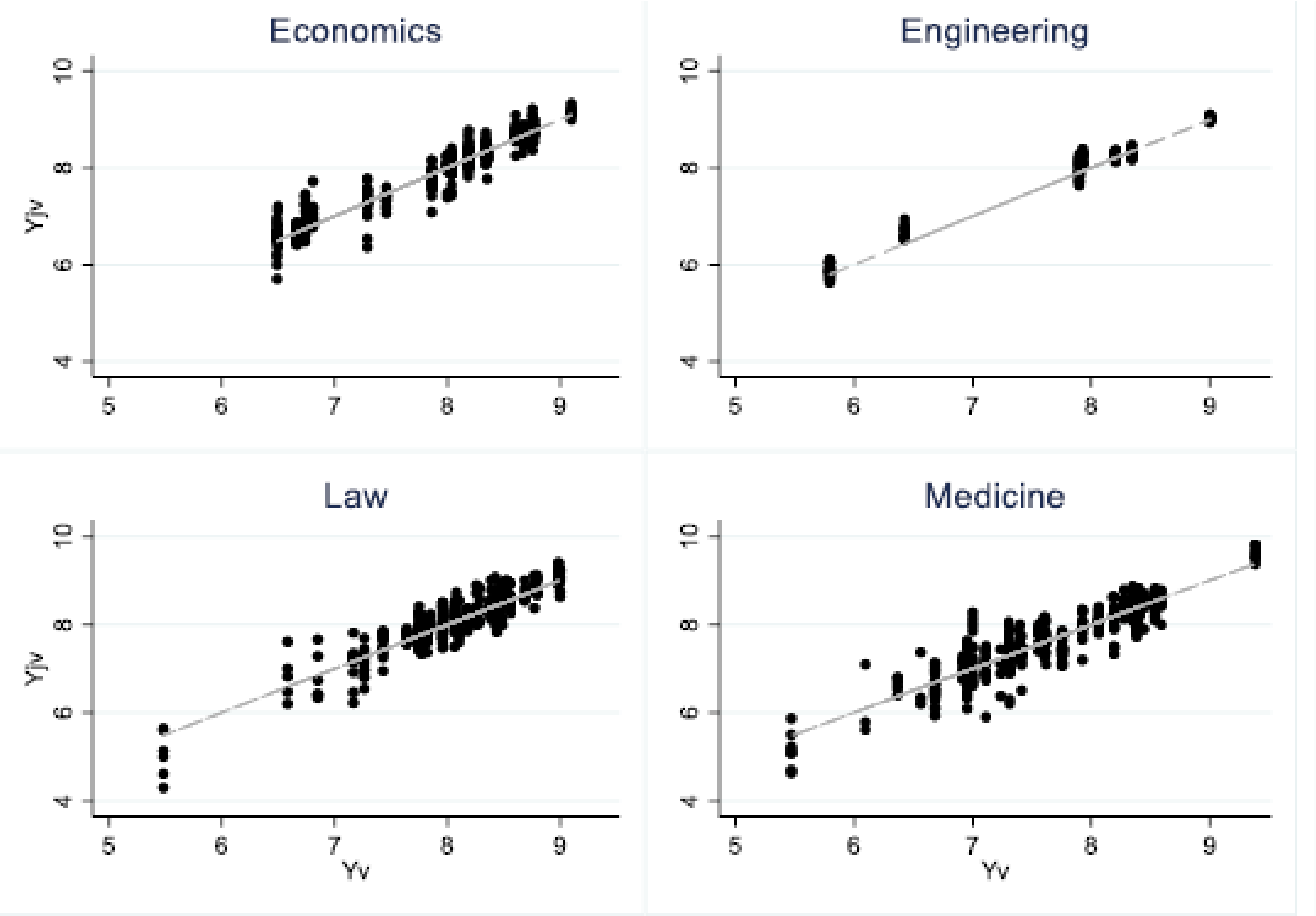
Testing for sorting on reporting styles: a solution

Trick: all students evaluate vignettes!

the **observed** average evaluation of a given vignette for **all** students	vs.	the **observed** average evaluation of that same vignette for the **subset of students** evaluating a specific elective

- Independent sorting: the distribution of reporting styles among the subsets of students evaluating each elective is the same as the one observed in the full population (up to sampling error)
- Testable implication: the average evaluation of a given vignette by the students of each subset should coincide with the average evaluation of the same vignette in the full population
- Conservative test given the total credits constraint imposed on students

Testing for sorting on reporting styles



Testing for sorting on reporting styles

Formal test: estimate $y_{jv} = \alpha + \beta y_v + \varepsilon_{jv}$ within major, test H0: $(\alpha = 0; \beta = 1)$

	Economics	Engineering	Law	Medicine
	(1)	(2)	(3)	(4)
α	0.188	0.340	-0.561	-0.149
	(0.095)	(0.095)	(0.187)	(0.168)
β	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: $\alpha = 0$	0.048	<0.001	0.003	0.377
H0: $\beta = 1$	0.092	0.008	0.001	0.369
H0: $(\alpha = 0; \beta = 1)$	0.002	<0.001	<0.001	0.664

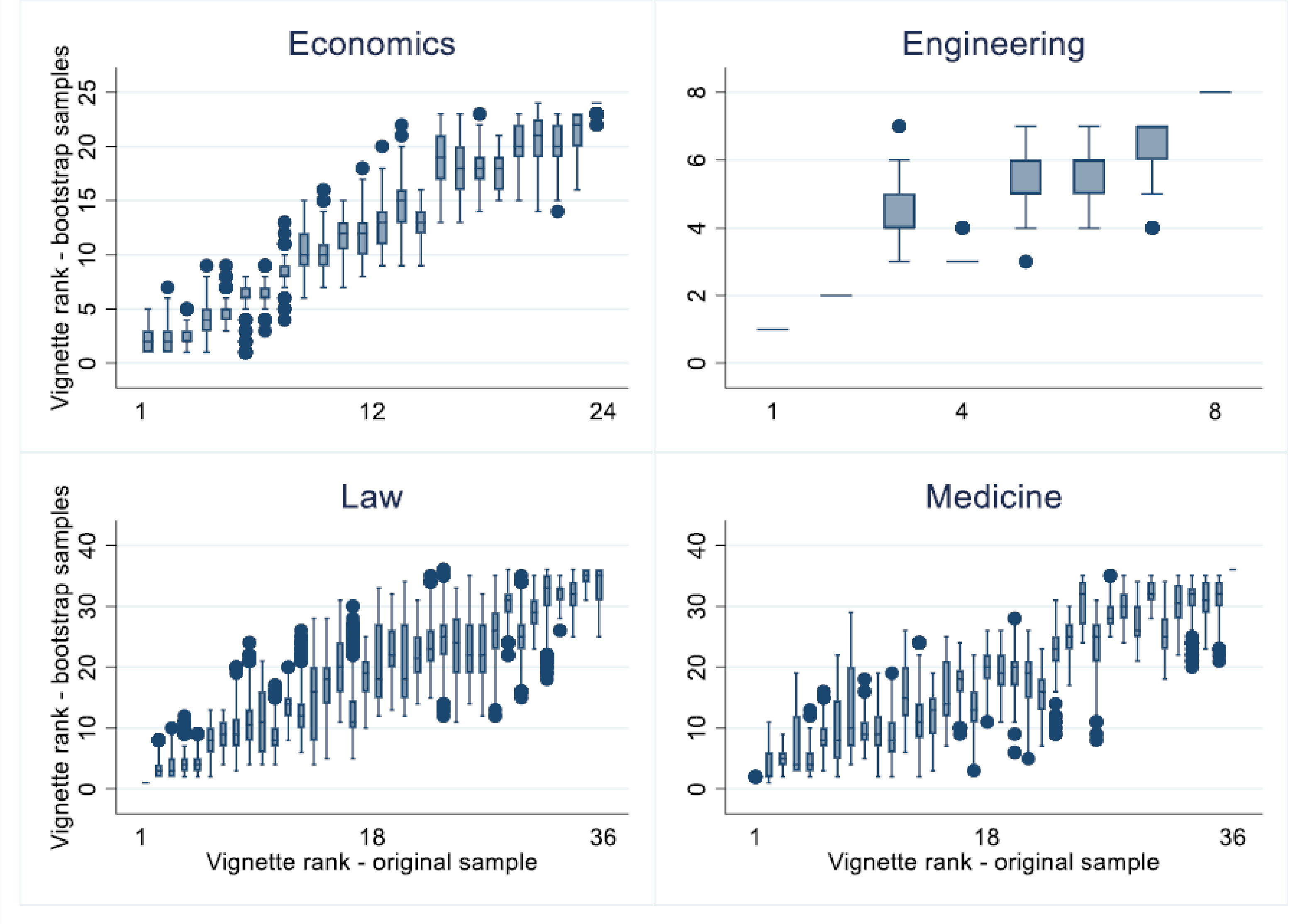
Consequences of sorting/noise for course ranking

Simulation exercise

- Draw at random one elective course per stratum
- Compute the average SET of each vignette for the sub-sample of student attending this elective
- Rank vignettes accordingly
- Redo 200 times

- This exercise illustrates the role played by sorting and noise to determine ranking of courses.

Consequences of sorting/noise for course ranking



Wrap up

- Clear evidence that SET ‘overall satisfaction’ is driven by:
    - Instructor ability to motivate
    - Instructor teaching clarity
    - (Student interest in the subject)

- We ask whether SET suffer from reporting heterogeneity

- We find that reporting heterogeneity accounts for between 25% and 45% of the within-course variability

- The ranking of a course may change significantly depending on the reporting styles/noise of the students who evaluates it.

Implications

1. SET do not provide a valid nor reliable estimate of course quality or teaching effectiveness
   - Stark and Freishtat (2014): pair them with evaluations of external experts

2. SET should not be used in comparative "tournament style" evaluations, because the ranking produced by SET can easily diverge from the ranking based on "true" course quality.

Is there room for improvement?

- SET can still be useful to evaluate teaching within a major if they are made comparable across students. How?
- Include specifically designed "vignette courses" in the curricula
  - courses of general content, comparable in all respects to other courses, that have to be attended and evaluated by all students at the beginning of their career
- Hardly feasible in practice
  - MOOC and online courses are a possibility, provided that response consistency holds
- Use SET to compare teaching across majors, departments, universities? Uhm…

MY TEACHER IS SHY AND WITHDRAWN, BUT I'M
SURE SHE'LL IMPROVE WITH TIME.