DISCUSSION PAPER SERIES

DP13883

ARE PROFESSORS WORTH IT? THE VALUE-ADDED AND COSTS OF TUTORIAL INSTRUCTORS

Jan Feld, Nicolás Salamanca and Ulf Zölitz

LABOUR ECONOMICS AND PUBLIC ECONOMICS

ISSN 0265-8003

ARE PROFESSORS WORTH IT? THE VALUE-ADDED AND COSTS OF TUTORIAL INSTRUCTORS

Jan Feld, Nicolás Salamanca and Ulf Zölitz

Discussion Paper DP13883
Published 22 July 2019
Submitted 16 July 2019

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

ARE PROFESSORS WORTH IT? THE VALUE-ADDED AND COSTS OF TUTORIAL INSTRUCTORS

Abstract

A substantial share of university instruction happens in tutorial sessions—small group instruction given parallel to lectures. In this paper, we study whether instructors with a higher academic rank teach tutorials more effectively in a setting where students are randomly assigned to tutorial groups. We find this to be largely not the case. Academic rank is unrelated to students' current and future performance and only weakly positively related to students' course evaluations. Building on these results, we discuss different staffing scenarios that show that universities can substantially reduce costs by increasingly relying on lower-ranked instructors for tutorial teaching.

JEL Classification: I21, I24, J24

Keywords: N/A

Jan Feld - jan.feld@vuw.ac.nz
Victoria University of Wellington

Nicolás Salamanca - n.salamanca@unimelb.edu.au
University of Melbourne

Ulf Zölitz - ulf.zoelitz@econ.uzh.ch
University of Zurich and CEPR

Are Professors Worth It?
The Value-added and Costs of Tutorial Instructors

Jan Feld Nicolás Salamanca Ulf Zölitz

Abstract

A substantial share of university instruction happens in tutorial sessions—small group instruction given parallel to lectures. In this paper, we study whether instructors with a higher academic rank teach tutorials more effectively in a setting where students are randomly assigned to tutorial groups. We find this to be largely not the case. Academic rank is unrelated to students' current and future performance and only weakly positively related to students' course evaluations. Building on these results, we discuss different staffing scenarios that show that universities can substantially reduce costs by increasingly relying on lower-ranked instructors for tutorial teaching.

Jan Feld is a Senior Lecturer at the School of Economics and Finance of the Victoria University of Wellington and IZA research affiliate. Nicolás Salamanca is a Research Fellow at the Melbourne Institute: Applied Economic & Social Research of The University of Melbourne and IZA research affiliate. Ulf Zölitz is an Assistant Professor at the Department of Economics of the University of Zurich, the Jacobs Center for Productive Youth Development, IZA research fellow CEPR and CESifo affiliate. *The authors thank Jeffrey Yusof, Sophia Wagner and Dominique Gilli for providing outstanding research assistance and Harold Cuffe, Alexandra de Gendre, Gabrielle Marconi, and participants in several seminars and conferences for useful comments and suggestions. This research was partly supported by the Australian Research Council Centre of Excellence for Children and Families over the Life Course (project number CE140100027). The Centre is administered by the Institute for Social Science Research at The University of Queensland, with nodes at The University of Western Australia, The University of Melbourne and The University of Sydney.* *Any correspondence about this paper should be directed to Ulf Zölitz, ulf.zoelitz@econ.uzh.ch.*

JEL classification: I21, I24, J24

I. Introduction

Instructors are a crucial, yet expensive input in university education. As a result, many universities have responded to cost pressures by increasingly relying on adjunct professors for lecturing.[1] A large share of university instruction, however, happens in tutorials. Tutorials—also called exercise, lab or TA sessions— are small group teaching sessions that cover material complementary to lectures. These tutorials are often responsible for more than half of the total wage costs per course, especially in large first-year courses.[2]

Universities differ widely in how they staff tutorials. In some universities, all tutorials are taught by students, but in others, tutorials are taught by a mixture of students and higher-ranked staff including full professors. The use of different types of instructors for tutorial teaching is surprising given the large differences in wage costs by academic rank. Professors, for example, cost much more than student instructors, which raises an obvious question: are professors worth it?

In this paper, we examine the costs and benefits of using tutorial instructors with different academic ranks. We use data from a Dutch business school where students within the same course are randomly assigned to instructors of different academic ranks, which range from fellow students to full professors. We first estimate individual instructors' value-added (VA) on course grades, grades in follow-on courses, course evaluations, as well as job satisfaction and earnings after graduation and then test whether each of these VA measures differ by instructor academic rank.

[1] See, for example, Ehrenberg (2012) concerning the increase of adjunct professors in the United States. Figlio, Schapiro, and Soter (2015) find that adjunct professors have a positive effect on student grades and that this effect is driven by low effectiveness of the bottom quarter of tenure track/tenured faculty. Bettinger and Long (2010) find that adjunct professors have a small positive effect on students' subsequent course enrollment.

[2] Because courses largely pair one large lecture with several tutorials, even moderate wage cost differences between lecturers and tutors will result in over half the courses' wage costs being caused by tutorials. To learn more about the prevalence of tutorial teaching, we conducted a small survey among OECD universities. The survey results suggest that 63 percent of OECD universities use tutorials, and in these universities, tutorials make up around 30 percent of students' contact hours. See Section II.A and Appendix 2 for a detailed description of the survey methodology and results. The survey data are available online at http://ulfzoelitz.com/research/material.

Based on these results, we then discuss the cost-savings potential of different staffing scenarios that rely on increasing the share of lower-ranked instructors.

We find that individual instructors significantly affect students' grades, follow-on grades, course evaluations, and job satisfaction, although effect sizes are quite small. An instructor with a one standard deviation higher VA increases students' grades by 2 percent of a standard deviation. Replacing the bottom 5 percent of instructors with average instructors would lead to a mere 0.3 percent of a standard deviation increase in the average student grade. Moreover, we cannot rule out the null hypothesis that individual instructors have no effect on earnings.

Instructors' academic rank is, overall, unrelated to students' academic outcomes. The most effective instructors—postdocs—add less than 1 percent of a standard deviation more to students' grades than student instructors. For all other instructor types, we can rule out differences between instructor types as small as 1 percent of a standard deviation in grades. Instructors' academic rank is also unrelated to students' grades in follow-on courses, where our results are also precisely estimated.

Looking at nonacademic outcomes, we find that instructors with higher academic rank add more value to students' course evaluations. However, these differences are also small. Students taught by a full professor, for example, evaluate the course only 4 percent of a standard deviation more positively than students taught by a student instructor. Finally, using matched survey data on university graduates, we find no systematic relationship between instructor academic rank and students' job satisfaction and earnings after graduation. These results are less precisely estimated, yet we can still rule out small-sized differences between most instructor ranks. Overall, our results suggest that replacing higher-ranked instructors with student instructors has no economically significant effects on students' current and future academic outcomes, job satisfaction and earnings

and only small negative effects on course evaluations. In other words, our results show little evidence that it is worth staffing tutorials with professors.

Building on these results, we conduct a simple accounting exercise that shows the savings potential under different tutorial staffing scenarios. In the most extreme scenario, in which all tutorials are taught by student instructors, wage costs for the average tutorial can be reduced by 49 percent for a bachelor's tutorial and by 52 percent for a master's tutorial. Under a more conservative scenario in which some potentially important higher-ranked instructors remain teaching in bachelor's tutorials and the staff composition for master's tutorials stays the same, we still calculate potential savings of 35 percent in bachelor's tutorials.

Previous studies have mostly focused on university instructors' effectiveness in lecturing large classes.[3] These studies consistently find that individual instructors matter. However, the estimated relationship between academic rank and instructor effectiveness differs substantially across studies. Carrell and West (2010) find that instructors at the U.S. Airforce Academy with a higher academic rank and terminal degree negatively affect students' current grades but positively affect students' future grades. Braga, Paccagnella, and Pellizzari (2016) find that instructors' academic rank at Bocconi University is unrelated to students' current grades, subsequent grades, and earnings after graduation. Hoffmann and Oreopoulos (2009) also find no significant relationship between academic rank and course dropout, grades, and course choice in a large Canadian University.[4] De Vlieger, Jacob, and Stange (2018) find that instructors' effectiveness in one algebra course at the University of Phoenix is unrelated to their salary.

[3] Another extensive strand of literature has looked at the effectiveness of teachers in primary and secondary education. This literature typically finds that teachers matter and are an important determinant of students' academic and labor market outcomes (for example, Chetty, Friedman, and Rockoff, (2014a)). However, there are conflicting findings about the relationship between formal qualifications and teacher effectiveness (for a review, see Harris and Sass (2011)).

[4] In another study, Bettinger, Long, and Taylor (2016) look at the effect of PhD student instructors compared to senior faculty on student course choice. They find that students are more likely to major in a subject if a PhD student taught the first course in that subject.

Only a few studies have looked at the effectiveness of instructors in tutorial teaching and other related tasks, such as holding office hours. These studies have focused on the ethnicity and origin of graduate teaching assistants (TAs). Lusher, Campbell, and Carrell (2018) study the role of graduate TAs' ethnicity and find that students' grades increase when they are assigned to same-ethnicity graduate TAs. Borjas (2000) and Fleisher, Hashimoto, and Weinberg (2002) study the effect of foreign-born as compared to native graduate TAs and reach opposing conclusions; Borjas (2000) finds that foreign-born TAs negatively affect student grades, whereas Fleisher, Hashimoto, and Weinberg (2002) find that foreign-born graduate TAs have negligible effects on student grades and that, in some circumstances, these effects can even be positive. None of these studies compare the effectiveness of tutorial instructors with different academic ranks.

We make three main contributions. First, this is the first study that focuses on instructors' effectiveness in tutorial teaching. Because tutorials are a critical part of many students' university education, our study fills an important knowledge gap in the literature on teacher effectiveness. Second, our rich data set allows us to look at a broad range of student outcomes, including course grades, student course evaluations, and various post-graduation labor market outcomes, giving us a comprehensive picture of instructors' impact on students in the short, medium, and longrun. Third, we discuss potential savings under different staffing scenarios. The size of these potential savings, along with our main results, will help university administrators make better-informed staffing decisions.

II. Background and Data

A. Tutorial Teaching in OECD Countries

To understand how common tutorial teaching is and how it differs between institutions, we

Table 1
Statistics on Tutorial Teaching in OECD Countries

	Obs.	Mean	Min	Max
Prevalence of tutorials:				
University uses small group (tutorial) teaching	69	0.63	0	1
Number of students scheduled per tutorial	49	22.23	0	140
Number of students attending per tutorial	49	16.34	2	100
All tutorial groups use the same course material	49	0.64	0	1
Programs that use tutorials:				
Only at the undergraduate level	49	0.32	0	1
Only at the graduate level	49	0.16	0	1
Both at the undergraduate and graduate level	49	0.53	0	1
Percentage of total contact hours spent in tutorials:				
Undergraduate	43	31.67	9	100
Graduate	23	26.89	9	50
Who teaches tutorials:				
Only students	49	0.25	0	1
Only professors	49	0.29	0	1
A mix of students & professors	49	0.46	0	1
Prevalence of instructor types in tutorials				
Bachelor's students	49	0.06	0	1
Master's students	49	0.25	0	1
PhD students	49	0.52	0	1
Teaching fellows	49	0.39	0	1
Adjunct instructors	49	0.23	0	1
Assistant professors	49	0.61	0	1
Associate professors	49	0.69	0	1
Full professors	49	0.64	0	1
Instructional method in undergraduate tutorials:				
Instructor stands in front of the class and explains	46	0.49	0	1
Instructor explains solutions to exercises	46	0.67	0	1
Students solve exercises	46	0.54	0	1
Students discuss material/exercise solutions	46	0.66	0	1
Students do group work	46	0.48	0	1
Instructional method in graduate tutorials:				
Instructor stands in front of the class and explains	23	0.28	0	1
Instructor explains solutions to exercises	23	0.44	0	1
Students solve exercises	23	0.60	0	1
Students discuss material/exercise solutions	23	0.86	0	1
Students do group work	23	0.57	0	1

Summary statistics representative for the OECD. Statistics calculated using poststratification weights by the share of universities in the country relative to the share of universities in the OECD. For more details, see Appendix 2.

conducted a small email survey among universities in OECD countries. In this survey, we gathered information about the nature of tutorial teaching from academic staff at 69 economics and business university departments in 31 OECD countries. We describe the survey questions and methodology in greater detail in Appendix 2.

We present some important insights from this survey in Table 1. In this table, we report weighted means to correct for oversampling of universities in small countries, using as weights the share of universities in the country relative to the share of universities in the OECD. The results indicate that 63 percent of universities in OECD countries offer tutorials at the undergraduate or graduate level. The average tutorial group size is 22 students. In universities where tutorials are used, students spend around 30 percent of their contact hours in tutorials. During these contact hours, students typically discuss and solve exercises, discuss course material, and do group work. Importantly, universities differ in how they staff tutorials. About 25 percent of all universities use only student or PhD-student instructors, whereas 46 percent use a mixture of student and higher-ranked instructors, such as assistant, associate, and full professors. About 29 percent of universities staff their tutorials exclusively with professors.

B. Institutional Environment and Sample Restrictions

To estimate the effect of instructor academic rank on student outcomes, we use data from a business school of a Dutch university for the academic years 2009–10 to 2014–15.[5] The bulk of the teaching at this business school is done in four regular teaching periods of eight weeks, during which students typically take two courses simultaneously. Note the distinction we make between "course" and "subject" throughout the paper: we use "subject" to refer to the material covered (for example,

[5] For more detailed information on the institutional environment, see Feld and Zölitz (2017) and Zölitz and Feld (2017).

Principles of Microeconomics) and "course" to refer to a subject-year-period combination (for example, Principles of Microeconomics in period 1 of 2011). Over the entire teaching period, students usually take three to seven 90-minute lectures for each course, which are taught by lecturers or assistant, associate, or full professors. The bulk of the teaching, however, is done in twelve two-hour tutorials. These tutorials are at the center of our analysis.

Tutorials are organized in groups of up to 16 students who are assigned to one instructor. In these tutorials, students discuss assigned readings or review solutions to exercises. As in many other universities, tutorials within a course are quite homogeneous: they use identical course material, have the same assigned readings and exercise questions, and follow the same course plan.[6] The main role of instructor is to guide tutorial sessions and help students when they are stuck. Instructors do not prepare their own lesson plan or select teaching material themselves, nor do they hold office hours. This narrowly specified role of the instructor allows us to isolate the effect of instructors' teaching delivery on student outcomes.

In contrast to other universities, tutorial attendance is compulsory, recorded by the instructor, and non-attendance can easily result in failing the course. Switching between assigned tutorial groups is explicitly prohibited and informal tutorial switching is extremely rare. The tutorial group composition we observe in our data is therefore near-identical to the assigned tutorial group composition.

Our institutional background and the method we use to calculate instructor VA impose some sample restrictions. Because we aim to distinguish course effects from instructor effects, we limit our sample to courses that were taught by at least two different instructors. Moreover, we follow Chetty, Friedman, and Rockoff (2014a) for our VA calculation (see Section III) in three

[6] The course material and plan are designed by the course coordinator (or, in rare instances, by two people who coordinate the course together), who is typically a lecturer, or an assistant, associate, or full professor. Course coordinators are not required to teach tutorials in the courses they coordinate, though often they do.

ways. First, we limit our estimation sample to instructors who teach the same subject in at least two periods because we need within-instructor time variation in outcomes to calculate our VA measures. Second, we exclude instructor-periods with fewer than seven students to make sure each VA estimate contains sufficient information. Third, because an instructor's effectiveness might differ substantially when he or she teaches different subjects, we consider each instructor-subject combination as a separate instructor. Thus, the same person teaching Microeconomics as well as Macroeconomics is counted as two separate instructors in our data. We discuss these and a few other sample restrictions in greater detail in Appendix 3. Our final estimation sample consists of 559 instructor-subject observations (which we refer to as instructors from now on), 651 different courses in 160 subjects, and 12,257 students.[7]

C. Tutorial Instructors

Students in the same course can be assigned to a student instructor, PhD student, postdoc, lecturer, assistant professor, associate professor, or full professor. Having different instructors teaching the same course allows us to separately identify the time-varying effects of individual instructors from the effect of course heterogeneity on student outcomes.

Table 2 describes our sampled instructors, the subjects they teach, and their wage costs by academic rank. Instructors' gender and nationalities vary widely across academic ranks, with lower shares of female and non-Dutch instructors in the higher academic ranks. Our data also reflect substantial variation by academic rank in the number of courses for which instructors taught tutorials, ranging from an average of 2.27 courses taught by student instructors to 3.68 taught by

[7] Table A1 in the Appendix 1 shows how the sample courses differ from the nonsample courses. Our sample courses are larger and rely more on students, PhD students, and lecturers as instructors. There is also a lower proportion of master's courses in the sample, as these are often taught by only one instructor. These differences do no lead to bias of our estimates because we do not use between-course variation in estimating instructor VA (see Section III). However, they are important for interpreting our results.

full professors. These differences mask even larger differences in teaching experience because higher-ranked instructors have accumulated more teaching experience before our sample period. Consistent with this, there are also differences in the number of tutorials and students taught across academic rank. PhD students are more likely to teach mathematical courses, whereas postdocs and lecturers teach more first-year courses.

Table 2
Summary Statistics by Instructor Academic Rank

	By instructor academic rank:						
	Student	PhD	Postdoc	Lecturer	Assist.	Assoc.	Prof.
Female instructor	0.53	0.34	0.43	0.33	0.22	0.17	0.15
Dutch instructor	0.24	0.19	0.18	0.60	0.39	0.62	0.87
German instructor	0.37	0.30	0.22	0.11	0.24	0.10	0.00
Belgian instructor	0.00	0.02	0.30	0.00	0.10	0.25	0.09
Other nationality instructor	0.15	0.47	0.30	0.28	0.27	0.04	0.00
Courses taught	2.27	2.70	3.23	3.54	3.56	3.23	3.68
Tutorials taught	7.18	7.31	9.17	10.80	8.84	7.85	9.04
Students taught	94.4	93.6	121.4	141.4	108.2	97.0	110.3
Mathematical courses taught	0.37	1.17	0.33	0.85	0.79	0.51	0.54
First-year courses taught	0.55	0.76	1.16	1.17	0.21	0.37	0.06
Hourly wage	€14	€20	€23	€31	€31	€43	€47
Total wage costs per tutorial session	€56	€79	€93	€126	€126	€173	€190
Instructors	55	157	20	182	85	32	28
Observations	3,942	12,254	1,882	19,763	6,735	2,235	2,031

This table summarizes demographic characteristics, teaching experience, course characteristics, and wages of instructors by their academic rank. The table is based on our estimation sample, which is comprised of 48,842 observations from 12,257 students who took 651 different courses in 160 different subject matters, taught by 559 instructors over 24 teaching periods between the academic years 2009-2010 and 2014-2015. Total wage costs per tutorial session include paid preparation time. Missing nationality information not shown.

Higher-ranked instructors also earn more than lower-ranked instructors. We base the wage costs reported in Table 2 on monthly gross wages from the lowest experience pay scale of the respective instructor rank (see Table A2 in the Appendix 1), giving us a lower bound of the actual wage costs of higher-ranked instructors. Still, the hourly wage of full professors, for example, is 3.4 times larger than the wage of student instructors, and twice as large as the wage of postdocs. The business school therefore pays €134 more for a tutorial session taught by a full professor than for a tutorial session taught by a student. These wage cost differences are themselves likely a lower bound for the total cost differences between instructor rank as they ignore overhead and hiring costs, which are usually greater for higher-ranked instructors. They are also not dampened by in-kind benefits received by lower-ranked instructors because student instructors do not receive tuition waivers or any other nonmonetary benefits besides their salary.

In addition to their different teaching responsibilities, instructors' contractual terms and nonteaching responsibilities differ as well. Professors, lecturers, and postdocs work on permanent or temporary contracts. Although lecturers mainly teach, professors and postdocs also perform academic research and fulfill administrative duties. PhD students and student instructors also pursue their own studies. PhD students are typically required to teach 20 percent of their contract time. Student instructors are often hired on short-term contracts when there are not enough regular staff or PhD students available to cover a course's teaching load; therefore, they disproportionally teach large bachelor's courses. They are typically hired by a university administrator mainly based on their grades, previous experience with the particular course, and a sufficient command of English, which is the language of instruction for all courses. Their contracts are always part-time and tutorial teaching is their only teaching obligation.

D. Student Outcomes

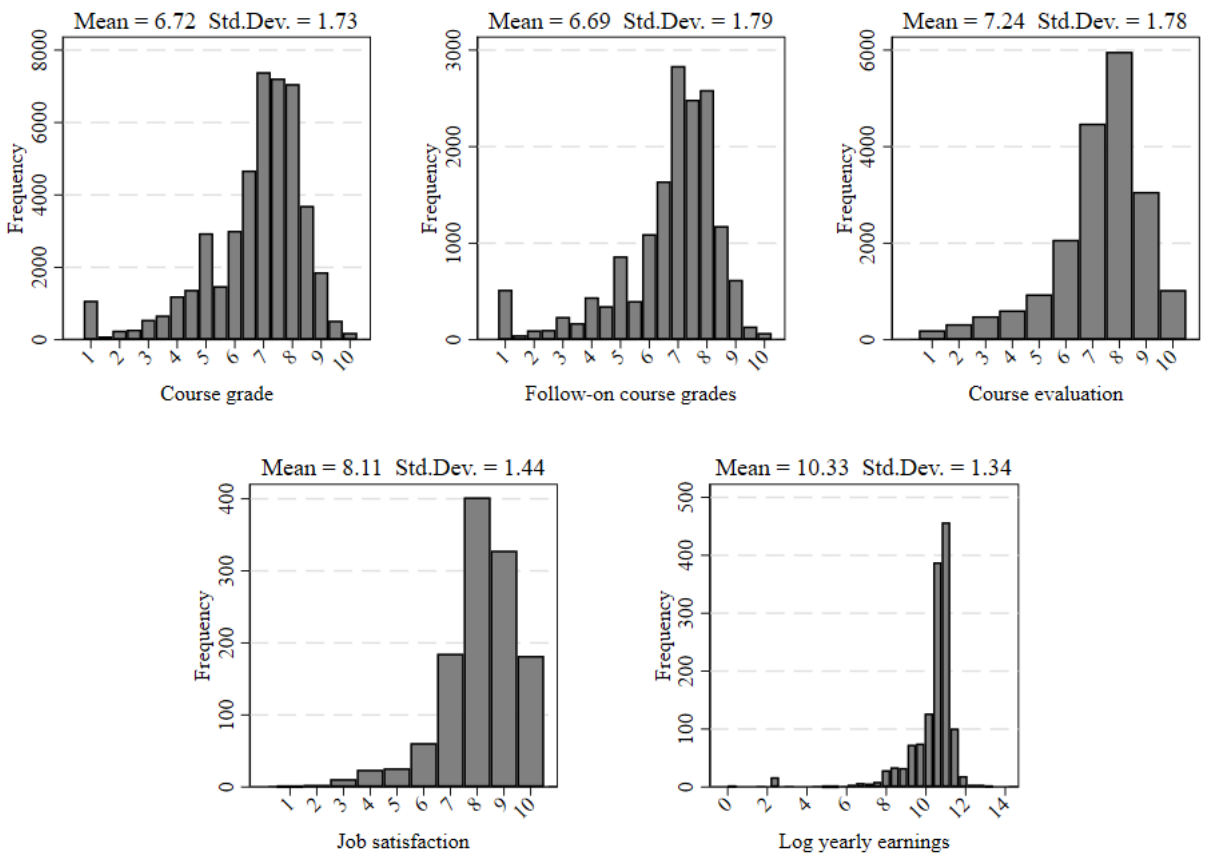
We estimate the effect of academic rank on five different student outcomes: course grades, grades in follow-on courses, students' course evaluations, and job satisfaction and earnings after graduation. Figure 1 shows the distributions of all these variables and reports their means and standard deviations in our estimation sample.

Course grades are given on a scale from 1 to 10, with 5.5 as the lowest passing grade. The final course grade usually consists of multiple graded components, with the highest weight typically placed on the final exam. All instructors involved in the course usually mark exams, with each instructor grading the same questions for all students. For example, for a 10-question exam in a course with two tutorial instructors, the first instructor may grade questions one through five, and the second instructor may grade questions six through ten for all students. Consequently, the student's own instructor only grades part of the exam.[8] The other graded components and their weights vary across courses, and some of these components, such as group work or tutorial participation, are graded directly by the students' instructor.

In our data we only observe final course grades, so one potential concern is that there are differences in grading standards by academic rank. We argue in Section IV that potential grading bias does not drive our results. Another potential concern is that we only observe grades of students who did not drop out of our sample. However, instructor rank is unrelated to course dropout, first-year completion, and on-time graduation (see Columns 1–3 of Table A3 in Appendix 1).

[8] See Feld, Salamanca, and Hamermesh (2016) for a detailed discussion of the examination and grading procedure.

Figure 1
Distribution of Student Outcomes



We define follow-on grades as the grades students receive in the next course on a similar subject matter, that is, the next course offered by the same department. There are eight different departments offering courses on a range of topics in economics, finance, and business.

We measure students' overall course evaluations with the following question included as part of the course evaluation surveys, which students take toward the end of the teaching period: *"Please give an overall grade for the quality of this course (1 = very bad, 6 = sufficient, 10 = very good)."* The course evaluation surveys have an average response rate of 42 percent, and there is some evidence of selective response related to instructor rank. The fourth column of Table A3 in the Appendix 1 shows that PhD students and full professors achieve significantly lower survey

response rates than student instructors. We show in Section IV that none of our results change once we account for this selectivity.

To collect data on students' job satisfaction and earnings after graduation, we surveyed students who obtained their undergraduate degree between 2011 and 2016. In this survey, we measure job satisfaction with the answer to the question: *"How satisfied are you, overall, with your current work?"*, which graduates could answer on a 10-point scale. We measure earnings with the answer to the question: *"What is your yearly income before taxes from your main job? (including bonuses and holiday allowances)."* The survey response rate was 37 percent, and in the matched data we have information on job satisfaction and earnings from 1,467 students in our estimation sample. The last column of Table A3 shows no evidence of selective response to this survey.[9]

E. Random Assignment of Students and Instructors to Tutorial Groups

A key feature of our setting is that, within a course, students are randomly assigned to tutorial groups conditional on scheduling conflicts.[10] For all bachelor students, this assignment was unconditionally random until the 2009–10 academic year. From 2010–11 onward, the scheduling office balanced tutorial groups by nationality (making sure that the proportion of German, Dutch, and other nationality students were the same across tutorial groups in each course), but otherwise, the assignment remained random. Instructors are then assigned to tutorial groups, generally in

[9] See Table A4 in the Appendix 1 for a comparison of the characteristics of our student population, estimation sample, and the respondent sample for the course evaluation and graduate surveys.

[10] Courses are usually scheduled in a way that avoids scheduling conflicts. For example, first-year compulsory courses that students take in parallel are scheduled on different days. The main source of scheduling conflicts is students' taking different elective courses. Accounting for potentially nonrandom assignment due to other courses taken at the same time by control for fixed effects for all combinations of courses that students take in each period leaves our results unchanged. A small number of students have other scheduling conflicts because they take language courses, work as student instructors, have regular medical appointments, or are top athletes and need to accommodate inflexible training schedules. Importantly, none of these exceptions is a response to the instructor or students of a tutorial group. One exception from the random assignment process is that before the fall of 2015, students could opt out of participating in tutorials that started at 6:30 p.m. Students in these evening tutorials represent only 6.6 percent of our observations. Limiting our estimation sample to courses without evening tutorials leads to qualitatively similar results.

consecutive time slots. Importantly, this assignment is unrelated to the characteristics of the students in the tutorial. About 10 percent of instructors in each period indicate time slots in which they are not available for teaching. However, this happens prior to any scheduling of students or other instructors and requires the approval of the department chair. Other papers using data from the same environment have shown that tutorial group assignment has the properties one would expect under random assignment (Feld and Zölitz, 2017; Zölitz and Feld, 2017; Mengel, Sauermann, and Zölitz, 2018).

Random assignment of students to tutorial groups implies that instructor characteristics are, in expectations, unrelated to observable and unobservable "pretreatment" student and tutorial group characteristics. To support this claim, we test whether in our estimation sample academic rank is related to: previous grade point average (GPA), gender, age, the rank of the student identification number (ID) (a proxy for tenure at the university), and tutorial size. We do this by regressing each of these five pretreatment characteristics on six instructor academic rank dummies (keeping student instructors as the base group), and instructor gender and nationality. We include course fixed effects as well as fixed effects for the tutorial sessions' time-of-day and day-of-the-week as controls.

Table 3 shows that academic rank is not systematically related to any of these five pretreatment characteristics. None of the F-tests for joint significance of the instructor rank dummies rejects the null hypothesis at the 5 percent level, nor do any of the F-tests for joint significance of all instructor characteristics. Looking at each academic rank coefficient individually, we see only that postdocs tend to teach younger students. However, this difference is small and likely due to chance given that we estimated 30 academic rank coefficients. Nevertheless, we include a cubic polynomial of student age, among other controls, when constructing VA measures. Overall, our results confirm that instructors' academic rank is not systematically related to pretreatment student and tutorial group characteristics.

III. Estimation of Instructor Value-added

A. Empirical Strategy

We estimate the effect of academic rank on instructor effectiveness by testing whether instructors' academic rank predicts their VA, that is, their independent contribution to a student's outcome. For simplicity, we focus our discussion in this section on current grades, but we also construct VA measures for other outcomes. Our VA construction broadly follows the methodology described in Chetty, Friedman, and Rockoff (2014a), which we implement using Michael Stepner's `vam` Stata program with minor modifications.

The VA construction process has three core steps. First, it models student grade as a function of student and tutorial group characteristics and then extracts the residuals from this model. These residuals contain the instructors' contributions to the student's grade and estimation noise. Second, it creates averages of the residuals at the instructor-time level. This step reduces the estimation noise, yet its main purpose is to aggregate data so that it varies at the appropriate level. Third, it predicts the average residuals for each instructor at each point in time with the average residuals of that same instructor *at every other point in time*. These predictions are the final VA measures.

Table 3
Balancing of Predetermined Characteristics on Academic Rank

Dep. Variable:	Previous GPA	Female	Age (years)	ID rank	Tutorial size
	(1)	(2)	(3)	(4)	(5)
Instructor academic rank (Base: Student)					
PhD	0.032	-0.021	-0.024	-90.470	0.213
	(0.043)	(0.017)	(0.053)	(133.151)	(0.174)
Postdoc	0.036	-0.005	-0.138**	-25.050	0.258
	(0.061)	(0.025)	(0.069)	(211.064)	(0.248)
Lecturer	0.029	-0.007	-0.004	-108.071	0.141
	(0.037)	(0.015)	(0.044)	(132.700)	(0.178)
Assist.	0.045	-0.014	-0.083	-180.238	0.071
	(0.043)	(0.018)	(0.054)	(151.702)	(0.188)
Assoc.	0.060	-0.026	0.032	-177.155	0.340
	(0.061)	(0.026)	(0.078)	(186.527)	(0.248)
Prof.	0.043	-0.011	-0.001	-146.954	0.005
	(0.066)	(0.028)	(0.091)	(225.900)	(0.247)
Instructor gender, nationality:	✓	✓	✓	✓	✓
Tutorial schedule FE:	✓	✓	✓	✓	✓
Course FE:	✓	✓	✓	✓	✓
F-test all inst. characteristics [p-value]	[0.927]	[0.871]	[0.129]	[0.977]	[0.675]
F-test inst. academic rank [p-value]	[0.961]	[0.874]	[0.066]	[0.922]	[0.456]
R-squared	0.70	0.05	0.49	0.01	0.74
Instructor-by-time	1,486	1,490	1,490	1,490	1,482
Observations	44,616	48,842	48,842	48,842	3,822

This table reports OLS coefficients of regressing student pre-determined characteristics on instructor characteristics. All regressions additionally include time-of-day and day-of-week fixed effects, and a dummy for students who registered late for the courses. Standard errors based on 500 pair bootstrap redraws clustered at the instructor-by-time level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1

The third step is the true innovation of the Chetty, Friedman, and Rockoff method. It adds three improvements to simply taking residual averages as VA measures. First, by predicting year-*t* residuals with all other year residuals, it makes sure that grade VA estimates do not include *contemporaneous* unexplained grades. This allows researchers to test whether grade VA predicts current grades without worrying about spurious results. Second, using all other years' averaged residuals as *separate* predictors allows residuals closer in time to be better predictors than residuals

that are further apart. Chetty, Friedman, and Rockoff show that this "drift adjustment" improves out-of-sample VA forecasts. Third, using predicted VA instead of residual averages shrinks VA estimates toward their mean. Because averaged residuals contain some estimation noise, which leads to biased estimates when using averaged residuals as regressors, the shrinkage ensures that the resulting VA measures are the "Best Linear Unbiased Predictor of a teacher's impact on average student [grades]" (Kane and Staiger, 2008, p. 2). Intuitively, shrinking average residuals toward the mean has the same effect as correcting regression coefficients for attenuation bias caused by measurement error.

To formally describe the steps in constructing our VA measures, we start by modeling the grade of student i taking course c and assigned to instructor j and time t as

$$grade_{icjt} = \beta' X_{ict} + \varepsilon_{icjt}\,, \tag{1}$$

where X_{ict} is a vector of comprehensive student and tutorial characteristics at time t. Student characteristics include a cubic polynomial of the student's age and dummies for gender, nationality, whether the student is in a bachelor's program, whether the student is an exchange student, whether the student is repeating the course, and whether the student is taking part in the business school's special research-based program. We also include a cubic polynomial of previous GPA, with all terms interacted with the repeat student dummy. Tutorial characteristics include tutorial size and tutorial-level averages of all student characteristics. We also include day-of-the-week and time-of-the-day fixed effects for the tutorials and a dummy for whether the student registered late for the course.[11] The parameter vector β captures the contributions of all these characteristics to the course grade, and we assume the error ε_{icjt} to have the following structure:

[11] We have a few missing values for nationality and age. We include a dummy for missing nationality, and we impute missing age as the tutorial-level mean or, if unavailable, the course-level mean. We create an imputed control dummy and interact it with all our controls in Equation (1). We impute a previous GPA of zero for the first period in our data,

(2) $$\varepsilon_{icjt} = \alpha_{jt} + \delta_{ct} + \nu_{icjt}.$$

In this error structure, time-varying course-specific unobserved heterogeneity, δ_{ct}, can be correlated with student and tutorial characteristics, X_{ict}, and with time-varying instructor-specific heterogeneity, α_{jt}. The correlation structure in ε_{icjt} captures obvious sorting patterns such as students sorting into courses based on their observable and unobservable characteristics as well as instructors systematically sorting into courses. Moreover, importantly, it allows both sorting patterns to be time-varying in an arbitrary way. ν_{icjt} is the usual random error term.

To construct our VA measures, we begin by estimating Equation (1) using a within-course transformation of the outcome and regressors, including a set of instructor fixed effects, and adjusting the variance-covariance matrix estimates from this regression to account for the additional parameters added by the within-course transformation (step 1). For the within transformation, we regress the transformed $\widetilde{grade}_{icjt} = grade_{icjt} - \overline{grade}_{ct}$ on $\tilde{X}_{icjt} = X_{icjt} - \bar{X}_{ct}$, where $\overline{grade}_{ct}$ and $\bar{X}_{ct}$ are course-level averages of the outcome and regressors in Equation (1). This is mathematically identical to adding course fixed effects in the estimation procedure, which the original `vam` program does not allow us to include because we are already absorbing instructor fixed effects. We add instructor fixed effects because leaving them out when estimating Equation (1) makes estimates of β understate the effect of student and tutorial characteristics if instructor VA is correlated with $\tilde{X}_{icjt}$, as we would otherwise attribute part of the instructor effect to these covariates (Chetty, Friedman, and Rockoff, 2014a). Therefore, by adding instructor fixed effects, we leave more variation in instructor effectiveness in the regression residuals, which are the basis for the VA estimates. More importantly, the within-course transformation eliminates the

where we cannot observe it, and interact a dummy for this period with our GPA measure. All these choices follow Chetty, Friedman, and Rockoff (2014a).

influence of δ_{ct} as a confounder. Our analysis exclusively relies on within-course variation, taking advantage of the random assignment of students to instructors and tutorial groups (see Section II.B). Estimating Equation (1) with only within-course variation justifies our identifying assumption that time-varying instructor unobserved heterogeneity—the key element of our instructor VA measure—is conditionally uncorrelated with other observable and unobservable determinants of grades. However, with this empirical design we cannot estimate differences in instructor VA across courses.

From the estimates of Equation (1) we construct the residuals:

$$grade^{*}_{icjt} = \widetilde{grades}_{icjt} - \hat{\beta}'\tilde{X}_{ict} = \hat{\alpha}_{jt} + \hat{v}_{icjt}. \tag{3}$$

These residuals are the basis for the grade VA estimates. We aggregate the residuals to instructor-time weighted averages, $\overline{grade}^{*}_{jt}$ using Chetty, Friedman, and Rockoff's precision weights, which give less weight to tutorial groups with more tutorial-level variance in grade residuals (step 2).

Finally, we predict the average residual grades at time t with average residual grades from all other times $k \neq t$ (step 3). Our final VA measure is equivalent to the predictions of averaged grade residuals at time t:

$$VA^{y}_{jt} = \sum_{k \neq t} \hat{\psi}_k \overline{grade}^{*}_{jk}\,, \tag{4}$$

where $\hat{\psi}_k$ is the bivariate OLS coefficient from regressing $\overline{grade}^{*}_{jt}$ on $\overline{grade}^{*}_{jk}$ for $k \neq t$, manually constructed using the corresponding autocovariances which efficiently uses all data. However, we restrict our residual autocovariances to be constant after $k > 4$, which is similar to imposing some equality restrictions on the OLS coefficients used for constructing predictions. This type of covariance restriction is illustrated clearly in Chetty, Friedman, and Rockoff (2014a). To enforce these restrictions, we manually generate the prediction coefficients by first estimating the time-

varying variances and autocovariances in $\overline{grade}^{*}_{jt}$ —the numerators and denominators of the restricted coefficients in instructor-year residual autoregressions—and then using them to construct the coefficients, $\hat{\psi}_k$, which is used for creating our final VA measures.[12]

Once we have calculated our VA estimates for our five different student outcomes, we follow Carrell and West (2010) by regressing all other VA estimates onto grade VA to explore the persistence of grade VA by estimating the following specification:

$$VA^{y}_{jt} = \gamma^{y} VA^{grade}_{jt} + \epsilon^{y}_{jt}\,, \tag{5}$$

where $y =$ follow-on grade, course evaluation, job satisfaction, and log-earnings.

Finally, we answer our main research question by testing whether our measures of instructor VA vary by instructor rank by estimating the following models:

$$VA^{y}_{jt} = \theta^{y} Rank_{jt} + e^{y}_{jt}\,, \tag{6}$$

where $Rank_{jt}$ is a vector of instructor rank dummies that excludes student instructors, which we leave as our base group. The coefficients of θ^{y} then show the differences in average VA between each instructor type and student instructors. We also explore the heterogeneous effects of instructor rank by estimating variations of Equation (6) for mathematical and non-mathematical subjects, and for first-year and non-first-year subjects. In these regressions, we cluster our standard errors at the instructor-by-time level to account for potential correlation of the error term within instructors. To increase efficiency, we produce all estimates of Equations (5) and (6) via weighted least squares, weighting each instructor-time observation by the square root of the number of students used to calculate its VA.

[12] Because the teaching in our setting is done in four regular teaching periods throughout the year, but most courses are taught on a yearly basis, our initial autocovariance estimates were sparse and had an inconvenient four-period cyclicality. To solve this issue, we restructured our data to have a synthetic instructor-specific time counter. We can do this without compromising the method's ability to account for common period shocks because we use within-course transformation in the construction of the residuals (step 1).

B. Estimates of Instructor Value-added

Table 4 summarizes our short- and longrun VA measures which, by construction, have a mean of zero. There is some variation in instructor VA, but it is quite compressed. One standard deviation of grade VA is 0.038, which means that an instructor who is one standard deviation more effective increases students' grades by 0.038 grade points on average, or 2 percent of a standard deviation in grades. To formally test whether the differences between instructors are statistically significant, we regress grade VA on instructor fixed effects and test whether these fixed effects are jointly significantly different from zero. Column (9) reports the p-value of this F-test, which shows that we have significant heterogeneity between instructors in grade VA.

Our estimated variation in instructors' grade VA in tutorial teaching is small compared to grade VA estimates in university lecturing, which range from 5 percent to 12 percent of a standard deviation (Braga, Paccagnella, and Pellizzari, 2016; Carrell and West, 2010; Hoffmann and Oreopoulos, 2009). As a reference, achievement VA estimates in primary and secondary school teaching range from 8 percent to 36 percent of a standard deviation (Hanushek and Rivkin, 2010).

Another way to illustrate the importance of instructors is to estimate the effect of replacing the bottom 5 percent of instructors with instructors of average quality (see Chetty, Friedman, and Rockoff 2014b; Hanushek, 2011). This back-of-the-envelope calculation is based on the distribution of instructors' average grade VA rather than the distribution of instructor-period grade VA displayed in Table 4. Replacing the bottom 5 percent of instructors with average grade VA instructors in this distribution would lead to a mere 0.3 percent of a standard deviation increase in the average student grade.[13]

[13] We do this back-of-the-envelope calculation in three steps. First, we calculate each instructor's mean grade VA, weighting by the square root of the number of students taught in each instructor-time period. Second, we create a counterfactual instructor mean VA distribution where we replace all instructors below the fifth percentile with the average instructor's VA. Third, we calculate the difference in means of both distributions, frequency-weighting by the total number of students taught by each instructor across all periods.

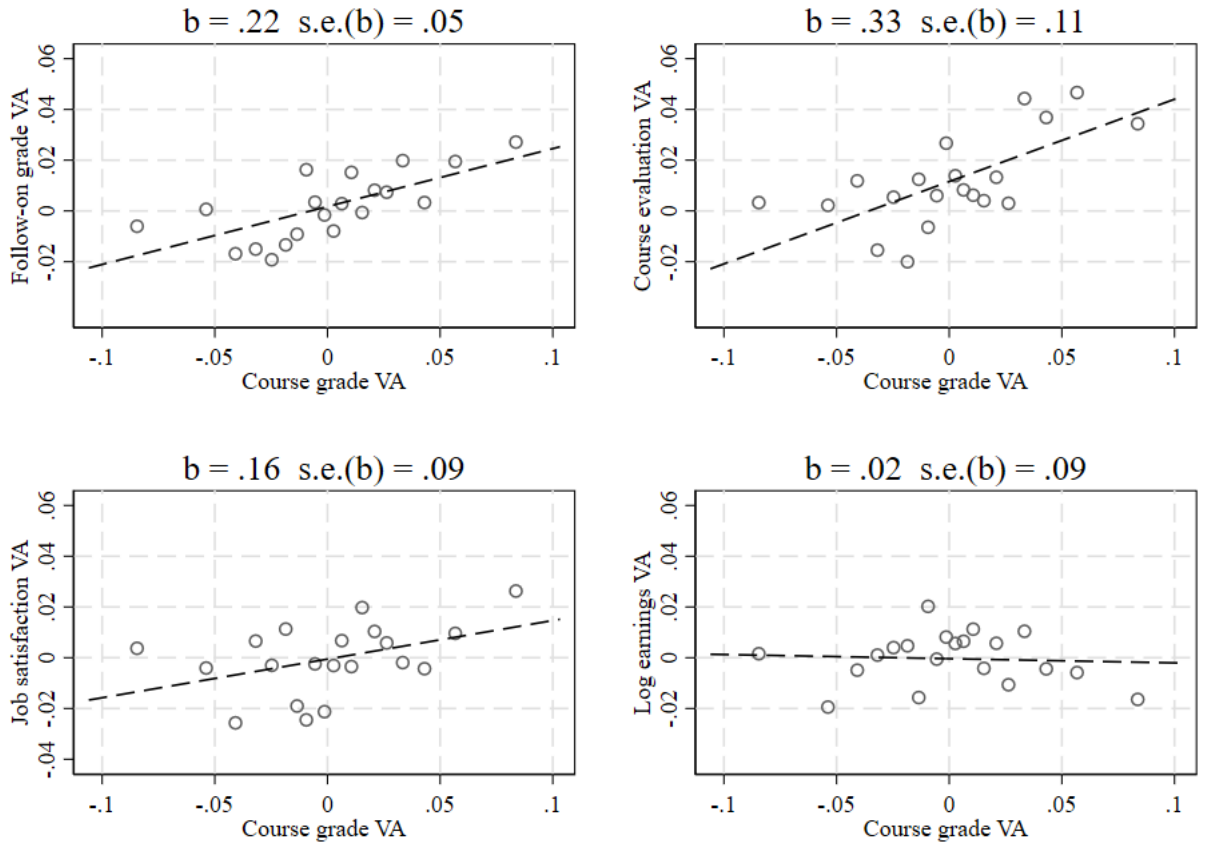
Our estimated standard deviation in follow-on grade VA is 0.054 grade points, which is quite similar in size to our grade VA estimates. The standard deviation of course evaluation VA, which, like grades, is measured on a 10-point scale, is about three times as large at 0.140 points. The standard deviation of job satisfaction VA is 0.091 points on a 10-point scale. For these three measures, we also observe significant heterogeneity between instructors. The standard deviation in log-earnings VA is quite large at 0.089, suggesting that a one-standard-deviation-more-effective instructor adds almost 9 percent to students' earnings. However, this large standard deviation is driven by the tails of the log-earnings VA distribution; its interquartile range is a much more modest 0.012. Also, for log-earnings VA we cannot reject the null hypothesis that individual instructors do not affect this labor market outcome.

Table 4
Summary Statistics of Value-added Estimates

				Percentile:					
	Obs.	No. inst.	Std. Dev.	1st	25th	50th	75th	99th	F-test of Instructor FE
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Value-added estimates:									
Course grade	1,432	502	0.038	-0.106	-0.021	0.001	0.023	0.102	[<0.000]
Follow-on grade	957	346	0.054	-0.152	-0.029	0.004	0.032	0.137	[<0.000]
Course evaluation	1,417	499	0.140	-0.318	-0.074	0.005	0.087	0.382	[<0.000]
Job satisfaction	1,299	478	0.091	-0.352	-0.032	0.002	0.036	0.298	[<0.000]
Log earnings	1,307	481	0.089	-0.384	-0.005	0.000	0.007	0.400	[>0.999]

This table reports summary statistics of value-added estimates at the instructor-period level for different outcomes. The number of observations differs by value-added measures due to missing values of outcomes. Column (9) reports the p-value of a joint significance test of the time-invariant instructor fixed effects as predictors of each value-added measure in square brackets, based on regular standard errors.

Figure 2
The Relation Between Different Instructor Value-added Estimates



These results naturally raise the question of how the different VA measures are related. Do instructors who raise students' contemporaneous grades also raise their future grades, their course evaluations, or their job satisfaction and earnings in the labor market? Figure 2 answers this question by showing scatterplots of the relation between grade VA and all other VA estimates by vingtiles of grade VA, as well as the bivariate regression coefficients underlying their linear relations in the disaggregated data. Instructors' effectiveness in raising students' current and future grades are significantly related. An instructor who adds one point to students' current grades also adds 0.22 points to their follow-on grades. This persistence of grade VA contrasts with Carrell and

West (2010), who find that current and future VA are negatively correlated. It is, however, consistent with Jacob, Lefgren, and Sims (2010), who also document persistence in primary and secondary education achievement VA. There is also a relationship between grade VA and course evaluation VA: instructors who add one more point to their student's grades get a course rating 0.33 points better. There is also some evidence that grade VA relates to job satisfaction VA, although the relation is half the size and only weakly statistically significant. Finally, there is no relation between grade VA and log-earnings VA.

IV. Academic Rank and Instructor Value-added

A. Main Results: Value-added by Academic Rank

Table 5 shows regression estimates of VA measures on dummies for academic rank, with student instructors as the base category. To ease the interpretation of our results, from this section onward we rescale our VA estimates by dividing them by the standard deviation of their respective outcome. Our regression coefficients then correspond to the VA differences between academic ranks in standard deviations of student outcomes. The exception is log-earnings VA, which we keep in log-points. For log-earnings VA, the regression coefficients should be interpreted as semi-elasticities, approximating percentage differences in earnings between each instructor type and student instructors.

Looking at all coefficients together, we find little evidence that students' course grades are systematically related to instructors' academic rank. While the F-test for joint significance rejects the null hypothesis that academic rank is unrelated to grade VA, all differences are economically tiny. The largest grade VA difference is between PhD students and postdocs, and it amounts to little more than 1 percent of a standard deviation in grades. These small differences though are

precisely estimated. We can, for example, rule out that full professors add more than 0.4 percent of a standard deviation in grades than student instructors.

Table 5
Value-added and Instructor Academic Rank

Dep. Variable:	*Value-added on:* Std. Course grade	Std. Follow-on grade	Std. Course evaluation	Std. Job satis-faction	Log earnings
	(1)	(2)	(3)	(4)	(5)
Instructor academic rank (Base: Student)					
PhD	-0.004	-0.007	-0.017**	0.003	0.001
	(0.002)	(0.004)	(0.008)	(0.008)	(0.007)
Postdoc	0.008**	0.003	0.046***	-0.017	0.007
	(0.003)	(0.006)	(0.015)	(0.012)	(0.014)
Lecturer	-0.003	-0.005	0.006	0.011	-0.003
	(0.002)	(0.004)	(0.008)	(0.008)	(0.010)
Assist.	0.005**	0.001	0.032***	0.020**	0.011
	(0.002)	(0.004)	(0.009)	(0.008)	(0.008)
Assoc.	0.003	0.002	0.044***	0.007	-0.007
	(0.003)	(0.005)	(0.011)	(0.010)	(0.011)
Prof.	-0.002	0.007	0.037***	0.011	0.010
	(0.003)	(0.006)	(0.011)	(0.009)	(0.009)
F-test inst. academic rank [p-value]	[<0.000]	[0.015]	[<0.000]	[0.002]	[0.117]
R-squared	0.02	0.01	0.06	0.01	0.00
Instructors	502	346	499	478	481
Observations	1,432	957	1,417	1,299	1,307

This table reports WLS coefficients of regressing measures of value-added on several student outcomes on instructor academic rank, weighting by the square root of the number of students identifying each value-added estimate. Heteroscedasticity-robust standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Follow-on course VA is also not related to academic rank. Our coefficients in this specification are again tiny and precisely estimated. This helps us dispel the concern that our grade VA estimates could reflect differences in grading standards on instructor-graded components or in

teaching to the test rather than student learning. Overall, we conclude that instructor academic rank is unrelated to students' current and future course performance.

Looking at the relationship between course evaluation VA and academic rank, we find that students evaluate a course more positively if they are taught by postdocs and professors of any rank. However, the estimated effect sizes are again small. The largest coefficient suggest that postdocs add 4.6 percent of a standard deviation to course evaluations over student instructors. PhD students lead to the worst course evaluations of all instructor types.

One concern is that these course evaluation VA estimates are driven by some small systematic differences in course evaluation survey responses by instructor rank. Compared to student instructors, PhD students and full professors achieve significantly lower response rates (see Table A3). This selective response may drive some of our results or hide even larger differences between instructor types, depending on what the course evaluation of the marginal nonresponding students would have been for each instructor rank. To correct for potential bias due to selective response, we follow Wooldridge (2007) and calculate course evaluation VA giving more weight to students who have a lower predicted probability of responding to the evaluation survey in a given course. These inverse probability weighted VA measures correlate almost perfectly ($\rho = 0.99$) with the original course evaluation VA measures. Unsurprisingly, our results are qualitatively identical when using the reweighted course evaluation VA measures as a dependent variable.[14]

Finally, we estimate instructor VA on job satisfaction and earnings measured in pretax log-points, both of which are important labor market outcomes. We thus test the possibility that

[14] Table A5 in the Appendix 1 shows the main results on course evaluations, job satisfaction, and earnings with inverse probability weighted (IPW) VA estimates to account for potentially selective survey response. The weights for the IPW analysis were calculated using predicted response probabilities from the model in the fourth column of Table A3 windsorized at the 1st and 99th percentiles. Under the assumption that our rich set of observed characteristics can inform the selection process (that is, the "coarsened at random" selection), the IPW estimator is consistent and more efficient than OLS (Wooldridge, 2007).

instructors affect students' labor market outcomes by, for example, giving career advice, even if they do not affect their grades. For job satisfaction VA, we do find evidence instructor rank matters, as shown by the significant F-test on instructor rank. The coefficients, however, indicate no systematic relationship between job satisfaction VA and academic rank. And the point estimates are again small: the largest coefficient suggests a mere 2 percent of a standard deviation difference between assistant professors and student instructors.[15]

For log-earnings VA, we again find no significant nor systematic heterogeneity between instructor types. The estimates, however, are somewhat noisier. This is particularly true for the effect of postdocs, where we would not be able to detect a wage premium smaller than 3.4 percent over student instructors. Differences for other instructor ranks are more precisely estimated; we can rule out log-earnings VA differences as small as 2.7 percent between professors and student instructors. On the other hand, even small increases in earnings would add up to significant amounts if we consider students' lifetime earnings or the effect of having more effective instructors in multiple courses. While we do not have enough statistical power to rule out any meaningful relationship between instructor rank and future earnings, we view our results as evidence that they are not strongly related.

Our main results support the idea that the academic rank of tutorial instructors does not relate to objectively-measured student outcomes. We can confidently rule out that academic rank is systematically related to students' current and future grades and largely conclude that instructor rank is unrelated to subsequent job satisfaction and earnings. Our results on course evaluation are suggestive of some nonpecuniary benefits of higher-ranked instructors, yet the magnitudes of these effects are minute. These results are consistent with the existing literature that has repeatedly shown

[15] Our estimates for job satisfaction VA and log-earnings VA are virtually identical when we use predicted survey response probabilities from the last column of Table A3 to obtain IPW VA estimates. We show these results in the second and third columns of Table A5 in Appendix 1.

that observable instructor characteristics are not strong determinants of differences in teacher VA (see Koedel, Mihaly, and Rockoff (2015) for a recent review).

B. Heterogeneity by Subject Type

Although we do not find meaningful differences in VA by academic rank, these average effects may still hide important heterogeneity by subject type. In this subsection we test whether higher-ranked instructors matter more for mathematical and non-first-year subjects, as these are presumably more difficult, which may affect the extent to which instructors can add value to their students. Moreover, looking at grade VA separately for first-year subjects provides us with a further test of whether grading biases are driving our main results. In first-year subjects, we expect grading biases to be smaller because for these subjects, the final grade we observe is equivalent to the exam grade, which often consists of machine-graded multiple-choice questions. Grading bias is likely to matter more in non-first-year subjects, which may contain components graded by the tutorial instructor like participation grades or presentation grades. If grading biases are hiding effects of academic rank on student learning, we would expect larger academic rank differences in grade VA for first-year subjects.

We define math subjects as those that have at least one of the following words in their description: *math*, *mathematics*, *mathematical*, *statistics*, *statistical*, *theory focused*. The definition of first-year subjects is self-explanatory. Empirically, we estimate the heterogeneity by subject type by regressing instructor VA on instructor rank dummies fully interacted with the indicators for subject type. We then estimate the average differences in VA across subgroups from these regressions.

Figure 3

Value-added and Instructor Academic Rank in Mathematical and Non-Mathematical Subjects

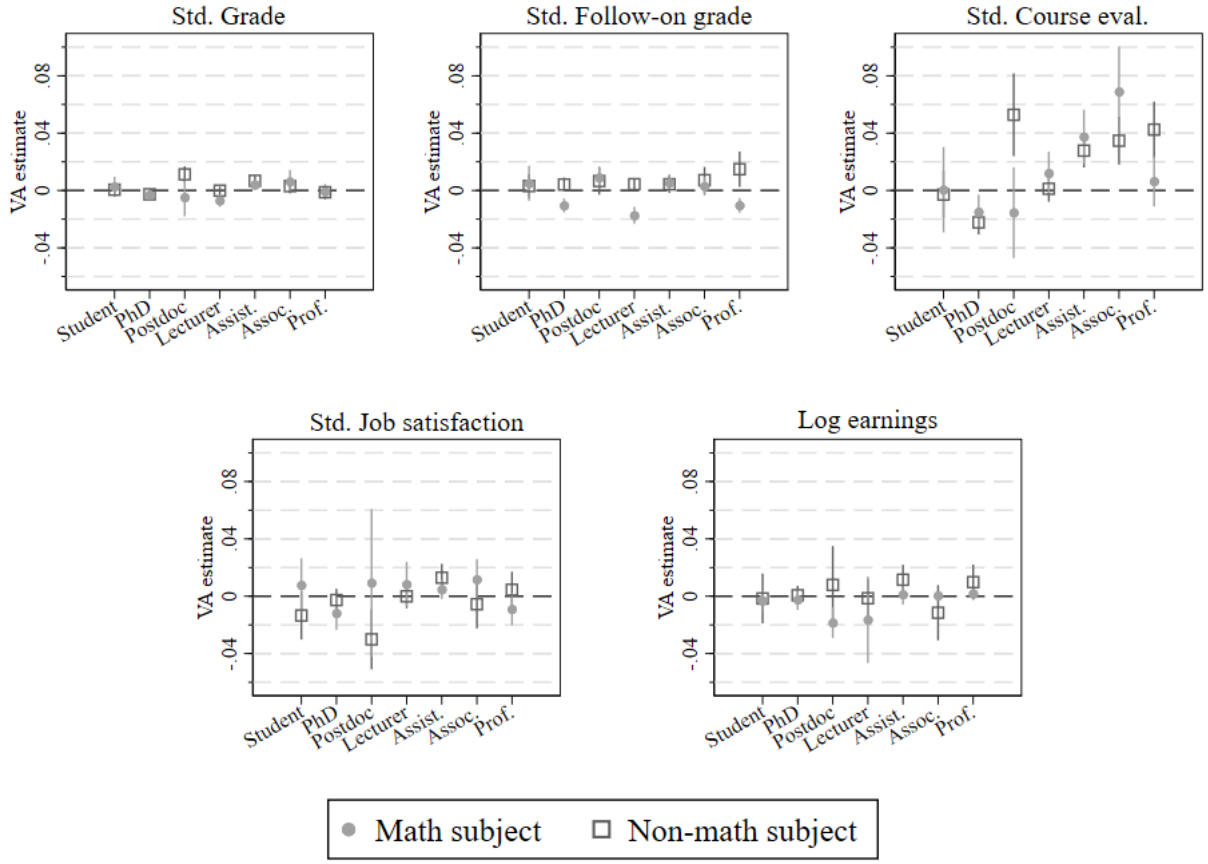


Figure 3 shows differences in VA for math and non-math subjects. For each of the five VA measures we see 14 estimates, which show the average VA of each of the seven academic ranks for math and non-math subjects. Looking at all 70 estimates together, we see no evidence that the effect of academic rank differs systematically between math and non-math subjects. The estimates are also small. When looking at grade and follow-on grade VA, we only see tiny point estimates—most of them smaller than 1 percent of a standard deviation. As in the main specification, the estimates for course evaluation VA are somewhat larger, but they also show no systematic difference in the performance of higher-ranked instructors between math and non-math subjects.

We also see no systematic or economically meaningful heterogeneity for job satisfaction VA or log-earnings VA. Overall, we find little evidence that the effect of academic rank on student outcomes differs by math course content.

Figure 4

Value-added and Instructor Academic Rank in First-Year and Non-First-Year Subjects

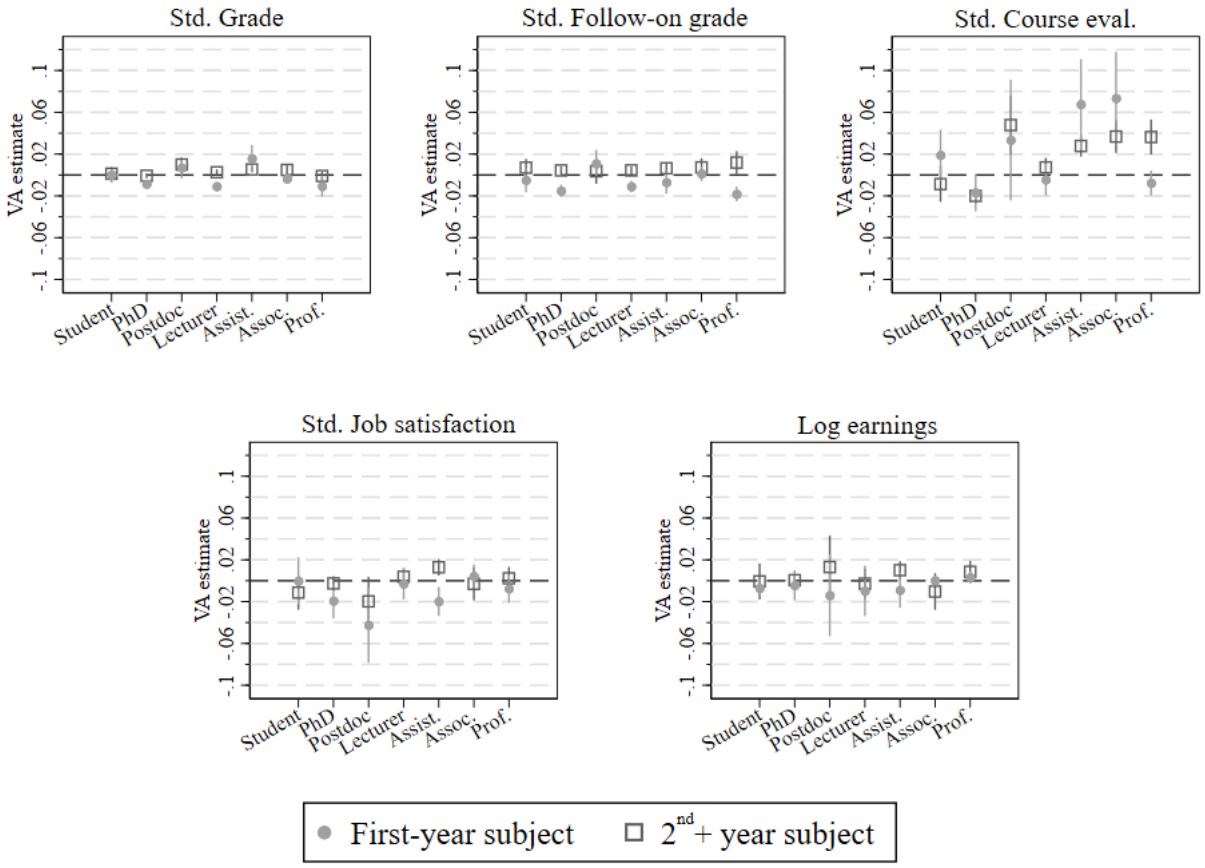


Figure 4 shows the average VA of instructors of different ranks for first-year and non-first-year subjects. The results again show no systematic or economically meaningful heterogeneity in any of the VA measures. In particular, academic rank differences in grade VA remain tiny and similar for both first-year and non-first-year courses, reinforcing the conclusion that grading biases

are not driving our main results. Differences in average follow-on grade VA, and job satisfaction VA are also economically insignificant, with all point estimates being smaller than 2 percent of a standard deviation. Although the average differences are again somewhat larger for course evaluation VA, there is no indication that higher-ranked instructors are systematically more or less effective in adding value in this dimension. Similarly, we see no systematic heterogeneity between first-year and non-first year subjects for job satisfaction VA and log-earnings VA. Taken together, our results suggest that there is little meaningful heterogeneity between first-year and non-first-year subjects.

V. Savings Potential Under Various Tutorial Staffing Scenarios

In the previous section, we have provided evidence that there are no economically meaningful benefits of having higher-ranked instructors in tutorials. Building on these results, we explore in this section the potential gains of using lower-ranked instructors. We do this by conducting an accounting exercise that showcases potential savings from changing the staff composition. These savings are driven by wage differences between instructor types at the business school. The magnitude of potential savings we discuss here is particularly informative for the 46 percent of OECD universities that also use a mixture of student and higher-ranked instructors. The savings potential at other institutions will of course depend on their current tutorial staffing arrangements and wage costs. This exercise nevertheless emphasizes the potential magnitude of two different cost-saving policies that could be implemented in many universities around the globe.

Table 6 shows the proportion of tutorials taught by instructors of different academic rank, and the average wage cost per tutorial in the status-quo. It also shows two alternative scenarios in which we replace higher-ranked with lower-ranked instructors. In the status-quo, lecturers and

professors of any rank teach 63 percent of all bachelor's and 62 percent of all master's tutorials. The wage cost per tutorial is thus €111 at the bachelor-level, and €117 at the master-level.

In our first alternative scenario, we calculate costs for the case in which student instructors would teach all bachelor's and master's tutorials. This scenario is similar to the situation of about 29 percent of all OECD universities, where students teach all tutorials. The average wage costs per tutorial in this scenario decreases to €56 for both bachelor's and master's tutorials. This is a 49 percent decrease in the wage costs for the average bachelor's tutorial and a 52 percent decrease for the average master's tutorial.[16]

These savings represent a substantial share of the business school's wage cost per course. Consider a typical five-lecture, five tutorial groups bachelor's course, where each tutorial group meets twelve times and all lectures are taught by an associate professor. The business school attributes eight hours for lecturing and preparation for each lecture, leading to wage costs of €1,720 for all five lectures. Given the average cost per bachelor's tutorial in the status-quo of €111, the wage costs for all 60 tutorial sessions would be €6,660 resulting in a total wage cost for the course of €8,380. Having all tutorials taught by student instructors would reduce the overall wage costs per course to €5,080, a 39 percent reduction.

It is worth pausing here to detail the conditions under which these savings can be realized without cost to the students of the business school. First, we are implicitly assuming that there are enough student instructors of similar quality as the ones sampled in our data to replace the higher-ranked instructors. This is akin to conducting our analysis in partial equilibrium. Second, we are leveraging our conclusion on the fact that we find no evidence that instructors of any rank have an

[16] Another alternative is to have PhD students teach all tutorials. Such a reassignment would lead to average wage costs per tutorial of €79, resulting in a 28 percent wage cost reduction for bachelor's tutorials and a 32 percent reduction for master's tutorials.

effect on later-life earnings. However, because each instructor affects around 33 students per period, even small earning penalties on student instructors would quickly accrue to large costs borne by students later on. Table 5 shows that we can reject the existence of quite small earning effects, which we explicitly interpret as evidence of no differences in earnings between instructors of different ranks. We lean on this interpretation for the current cost exercise.

Table 6
Tutorial Wage Costs Under Different Staffing Scenarios

		Status quo		*Scenario 1*		*Scenario 2*	
		Bachelor	Master	Bachelor	Master	Bachelor	Master
	Wage costs per tutorial session	Percentage of tutorials currently taught by…		Student instructor teaching all tutorials		Keep course coordinators and PhD students	Staff composition unchanged
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Student	€56	9%	5%	100%	100%	64%	5%
PhD	€79	24%	29%	0%	0%	24%	29%
Postdoc	€93	4%	4%	0%	0%	0%	4%
Lecturer	€126	46%	22%	0%	0%	5%	22%
Assist.	€126	10%	25%	0%	0%	4%	25%
Assoc.	€173	4%	6%	0%	0%	2%	6%
Prof.	€190	3%	9%	0%	0%	1%	9%
Average wage costs per tutorial		€111	€117	€56	€56	€72	€117
Total savings potential				49%	52%	35%	0%

This table reports wage costs per tutorial and share of staff allocated to tutorials by instructor ranks. For the average wage costs per tutorial, gross wages are assumed to be in the lowest pay scale of the instructor type. This assumption leads to a more conservative estimate of the savings potential because the actual cost reduction for substituting senior instructors is underestimated.

Finally, we are not monetizing the small effect on course evaluations of replacing higher-ranked instructors by student instructors. Course evaluations may have a monetary value to the business school; moreover, the decrease in course evaluation ratings could be reflecting a loss of

nonpecuniary value of education for the students. However, course evaluations would have to be immensely valuable to the business school to warrant the use of the more expensive faculty as tutorial instructors. For example, hiring a full professor instead of a student instructor for one tutorial group and a set of twelve tutorial sessions would cost the business school an additional €1,608 in wages. The expected return on this investment would be a 3.7 percent of a standard deviation increase in course evaluations. The break-even point of this trade-off would imply a valuation of at least €43,460 per standard deviation increase in course evaluations per tutorial group for the business school. Increasing course evaluations by hiring postdocs would be a more cost-effective, though still very expensive, way to increase course evaluations with an implied valuation of at least €9,652 per standard deviation increase in course evaluations per tutorial group.

There are at least three reasons for considering a less extreme scenario in which higher-ranked instructors still teach some tutorials. First, having the course coordinator teach at least one tutorial may allow them to adjust the content of the lectures, adapt the learning material or exam content, and give advice to lower-ranked instructors. Empirically, we do not identify these spillovers that would benefit all students in a course because our estimates shown in Section IV only use within-course variation in the VA construction. Second, taking PhD students out of the teaching force might have unintended negative consequence for their job prospects, especially in academia (see Bettinger, Long, and Taylor, 2016). Third, our VA estimates are mainly driven by bachelor's courses and our results may not generalize to all master's courses. For example, many of the smaller master's courses that are excluded from our estimation sample because they only had one instructor, may be too technical for student instructors to teach.

In our second alternative scenario, we keep these caveats in mind and simulate a counterfactual staff assignment in which we do not change the staff composition in master's courses, keep the status-quo share of PhD students in bachelor's courses, and allow the highest-

ranked instructors in each bachelor's course to teach one tutorial. In this scenario, the average wage costs decrease from €111 to €72 for bachelor-level tutorials—a 35 percent reduction compared to the status-quo. This reduction, although smaller than in our first counterfactual scenario, still signifies a large cut in wage costs for the business school.

Universities should, of course, do more than an accounting exercise before changing their staffing policies. In many situations, it may not be feasible or desirable to dramatically change the staff composition, especially in the short run. However, in all cases, universities should still consider the opportunity costs of time for higher-ranked instructors. These opportunity costs likely differ between instructors. For example, there might be research-inactive and tenured professors for whom teaching tutorials would be the most valuable use of time. However, we generally believe that professors are more valuable doing other activities such as research. Although there are a number of factors to consider, many idiosyncratic to the specific institution, we believe that increasingly relying on lower-ranked instructors is a promising avenue to explore for universities that seek to reduce costs or want to give their professors more time for research.

VI. Conclusion

Universities around the world have very different policies on how they staff small teaching sessions, often referred to as tutorials. In this paper, we investigate how tutorial instructors' academic rank relates to their teaching effectiveness as measured by how much value they add to students' course grades, students' grades in follow-on courses, the evaluations students give to the courses, and students' subsequent job satisfaction and earnings. We show that, despite substantial differences in formal qualifications and wage costs, instructor academic rank is unrelated to students' current and follow-on grades. Put differently, professors are not better than student instructors in increasing student performance. Our estimates are precise enough to rule out very

small differences in instructor performance. For example, we can rule out differences as large as 1 percent of a grade standard deviation in teaching effectiveness between full professors and students. We find evidence that professors receive marginally higher course evaluations. Yet, these estimates are economically miniscule. We find no evidence that academic rank is systematically related to students' job satisfaction or earnings.

There might be, of course, differences in teaching effectiveness that we do not capture with our broad range of outcomes. For example, professors might be better at dealing with students' family problems and mental health issues. Lower-ranked instructors could also offer benefits to their students that we do not observe. For example, student and PhD student instructors might be better able to give students informal advice on how to study for exams and which elective courses to take. We doubt that these unobserved differences would justify the substantially higher cost of staffing tutorials with higher-ranked instructors.

As with other studies that rely on data from one institution, it is not clear how our results translate to other contexts. Tutorials in other universities could be intrinsically different in ways that challenge the external validity of our findings. For example, at the business school we study, the main role of the instructor is to guide classroom discussions. Academic rank may matter more in settings where the instructors' main role is to explain the course material. This is an important empirical question.

Taken together, our results raise an important question: Is tutorial teaching really the best use of a professor's time? Our findings suggest that universities could increase research output by assigning students instead of professors to tutorial teaching. We suspect most professors would be fine with that.

References:

Bettinger, Eric P., Bridget Terry Long, and Eric S. Taylor. 2016. "When Inputs Are Outputs: The Case of Graduate Student Instructors." *Economics of Education Review* 52:63–76. https://doi.org/10.1016/j.econedurev.2016.01.005.

Bettinger, Eric P, and Bridget Terry Long. 2010. "Does Cheaper Mean Better? The Impact of Using Adjunct Instructors on Student Outcomes." *Review of Economics and Statistics* 92 (3):598–613. https://doi.org/10.1162/REST_a_00014.

Borjas, George J. 2000. "Foreing-Born Teaching Assistants and the Academic Performance of Undergraduates." *American Economic Review* 90 (2):355–59. http://www.jstor.org/stable/117250.

Braga, Michela, Marco Paccagnella, and Michele Pellizzari. 2016. "The Impact of College Teaching on Students ' Academic and Labor Market Outcomes." *Journal of Labor Economics* 34 (3):781–822. https://doi.org/https://doi.org/10.1086/684952.

Carrell, Scott E., and James E. West. 2010. "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors." *Journal of Political Economy* 118 (3):409–32. https://doi.org/10.1086/653808.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014a. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review* 104 (9):2593–2632. https://doi.org/10.1257/aer.104.9.2593.

———. 2014b. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review* 104 (9):2633–79. https://doi.org/10.1257/aer.104.9.2633.

Ehrenberg, Ronald G. 2012. "American Higher Education in Transition." *Journal of Economic Perspectives* 26 (1):193–216. https://doi.org/10.1257/jep.26.1.193.

Feld, Jan, Nicolás Salamanca, and Daniel S. Hamermesh. 2016. "Endophilia or Exophobia: Beyond Discrimination." *The Economic Journal* 126 (594):1503–27. https://doi.org/10.1111/ecoj.12289.

Feld, Jan, and Ulf Zölitz. 2017. "Understanding Peer Effects: On the Nature, Estimation, and

Channels of Peer Effects." *Journal of Labor Economics* 35 (2):387–428. https://doi.org/10.1086/689472.

Figlio, David N., Morton O. Schapiro, and Kevin B. Soter. 2015. "Are Tenure Track Professors Better Teachers?" *Review of Economics and Statistics* 97 (4):715–24. https://doi.org/10.1162/REST_a_00529.

Fleisher, Belton, Masanori Hashimoto, and Bruce A. Weinberg. 2002. "Foreign GTAs Can Be Effective Teachers of Economics." *Journal of Economic Education* 33 (4):299–325. https://doi.org/10.1080/00220480209595329.

Hanushek, Eric A. 2011. "The Economic Value of Higher Teacher Quality." *Economics of Education Review* 30 (3):466–79. https://doi.org/10.1016/J.ECONEDUREV.2010.12.006.

Hanushek, Eric A., and Steven G. Rivkin. 2010. "Generalizations about Using Value-Added Measures of Teacher Quality." *American Economic Review* 100 (2):267–71. https://doi.org/10.1257/aer.100.2.267.

Harris, Douglas. N., & Sass, Tim. R. (2011). Teacher Training, Teacher Quality and Student Achievement. *Journal of Public Economics*, 95(7-8), 798-812. https://doi.org/10.1016/j.jpubeco.2010.11.009

Hoffmann, Florian, and Philip Oreopoulos. 2009. "Professor Qualities and Student Achievement." *Review of Economics and Statistics* 91 (1):83–92. https://doi.org/10.1162/rest.91.1.83.

Jacob, Brian A, Lars Lefgren, and David P. Sims. 2010. "The Persistence of Teacher-Induced Learning." *Journal of Human Resources* 45 (4):915–43. https://doi.org/10.1353/jhr.2010.0029.

Kane, Thomas, and Douglas Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." *National Bureau of Economic Research*. Cambridge, MA. https://doi.org/10.3386/w14607.

Koedel, Cory, Kata Mihaly, and Jonah E. Rockoff. 2015. "Value-Added Modeling: A Review." *Economics of Education Review* 47 (August):180–95. https://doi.org/10.1016/j.econedurev.2015.01.006.

Lusher, Lester, Doug Campbell, and Scott Carrell. 2018. “TAs like Me: Racial Interactions between Graduate Teaching Assistants and Undergraduates.” *Journal of Public Economics* 159 (March):203–24. https://doi.org/10.1016/j.jpubeco.2018.02.005.

Mengel, Friederike, Jan Sauermann, and Ulf Zölitz. forthcoming. “Gender Bias in Teaching Evaluations.” *Journal of the European Economic Association*. https://doi.org/10.1093/jeea/jvx057.

Vlieger, Pieter De, Brian Jacob, and Kevin Stange. 2018. “Measuring Effectiveness in Higher Education.” In *Productivity in Higher Education*, ed. by Caroline M. Hoxby and Kevin Stange. University of Chicago Press. http://www.nber.org/chapters/c13880.pdf.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, 141(2), 1281-1301. https://doi.org/10.1016/j.jeconom.2007.02.002

Zölitz, Ulf, and Jan Feld. 2017. “The Effect of Peer Gender on Major Choice.” *University of Zurich, Department of Economics Working Paper Series*. https://doi.org/10.2139/ssrn.3071681.

Appendix 1
Additional Tables

Table A1
Comparison of Sample vs. Nonsample Courses

	Sample courses (N = 651) Mean	Other courses (N = 628) Mean	Diff. in Means
	(1)	(2)	(2) - (1)
Instructor academic rank:			
Student	0.18	0.12	0.06
PhD	0.27	0.14	0.13
Postdoc	0.03	0.05	-0.02
Lecturer	0.28	0.14	0.14
Assist.	0.14	0.25	-0.11
Assoc.	0.05	0.18	-0.13
Prof.	0.05	0.12	-0.07
Student characteristics:			
Grade	6.85	7.1	-0.25
Previous GPA	6.11	6.31	-0.20
Bachelor	0.65	0.46	0.19
Course characteristics:			
Mathematical	0.27	0.42	-0.15
First-year	0.18	0.1	0.08
Offered by microeconomics dept.	0.12	0.16	-0.04
Offered by macroeconomics dept.	0.06	0.12	-0.06
Offered by finance dept.	0.16	0.08	0.08
Offered by other dept.	0.66	0.65	0.01
No. instructors	4.01	1.18	2.83
No. students	140.71	31.66	109.05
No. tutorials	10.88	2.71	8.17
No. students per tutorial	12.55	11.24	1.31

This table is based on data from 111,481 observations from 14,051 students who took 1,279 courses in 160 different subject matters, taught by 2,054 instructors over 24 teaching periods between the academic years 2009-2010 and 2014-2015.

Table A2
Wage Costs and Contractual Time by Instructor Academic Rank

	By instructor academic rank:						
	Student	PhD	Postdoc	Lecturer	Assist.	Assoc.	Prof.
Monthly gross wage	€2,251	€3,179	€3,714	€5,034	€5,034	€6,908	€7,599
FTE teaching and preparation (hours per month)	*flexible*	32	40	160	80	80	80
FTE Standard teaching load	*flexible*	0.20	0.25	1	0.50	0.50	0.50
Hourly wage	€14	€20	€23	€31	€31	€43	€47
Hours per tutorial session in:							
Paid preparation	2	2	2	2	2	2	2
Teaching	2	2	2	2	2	2	2
Total	4	4	4	4	4	4	4
Total wage costs per tutorial session	€56	€79	€93	€126	€126	€173	€190

Monthly gross wages are assumed to be in the lowest pay scale of the instructor type, which provides a lower bound of the actual costs for more senior instructors. Calculations based on a total of 160 Full Time Equivalent (FTE) hours in a month.

Table A3

Dropout, Course Evaluation Response, and Survey Response by Instructor Academic Rank

Dep. Variable:	Course dropout	First-year dropout	On-time graduation	Course eval. respondent	Survey respondent
	(1)	(2)	(3)	(4)	(5)
Instructor academic rank (Base: Student)					
PhD	0.002	0.010	-0.008	-0.032*	0.012
	(0.009)	(0.009)	(0.022)	(0.019)	(0.016)
Postdoc	-0.004	-0.005	0.019	-0.041	-0.012
	(0.011)	(0.012)	(0.036)	(0.034)	(0.023)
Lecturer	-0.003	0.001	0.002	-0.006	-0.003
	(0.008)	(0.007)	(0.021)	(0.017)	(0.014)
Assist.	-0.000	0.003	0.003	-0.016	-0.012
	(0.009)	(0.008)	(0.026)	(0.020)	(0.016)
Assoc.	-0.002	0.004	0.007	-0.025	-0.002
	(0.011)	(0.010)	(0.038)	(0.024)	(0.019)
Prof.	-0.002	0.001	0.002	-0.061**	-0.016
	(0.012)	(0.009)	(0.042)	(0.027)	(0.023)
Instructor gender, nationality, experience:	✓	✓	✓	✓	✓
Tutorial schedule FE:	✓	✓	✓	✓	✓
Course FE:	✓	✓	✓	✓	✓
F-test inst. academic rank [p-value]	[0.987]	[0.754]	[0.987]	[0.171]	[0.466]
R-squared	0.07	0.67	0.08	0.09	0.12
Instructors-by-time	1,490	1,015	907	1,490	1,433
Observations	48,842	34,350	24,236	48,842	41,390

This table reports OLS coefficients of regressing student dropout and survey response dummies on instructor observable characteristics. All regressions condition time-of-day and day-of-week fixed effects, a dummy for students who registered late for the courses, and course fixed effects. Standard errors based on 500 pair bootstrap redraws clustered at the instructor-by-time level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table A4

Comparison of Student Characteristics for the Population and the Different Estimation Samples

	Student population:		*Student sample:*		
	All	Graduate Survey	Estimation	Course evaluation	Graduate Survey
	N = 14,051	N = 10,566	N = 12,257	N = 7,574	N = 1,737
	(1)	(2)	(3)	(4)	(5)
Female	0.40	0.40	0.40	0.44	0.38
Age	20.71	21.00	20.67	20.72	19.92
Dutch	0.25	0.26	0.26	0.25	0.32
German	0.36	0.37	0.38	0.41	0.57
Course grade	6.76	6.77	6.72	6.86	7.04
Total courses taken	7.93	9.29	3.98	2.48	7.71

This table reports means of student characteristics for the population of students in our data (comprising 111,481 observations from 14,051 students who took 1,271 different courses in 274 subject matters, taught by 2,054 instructors over 24 teaching periods between 2009 and 2014), the population of students eligible to answer the graduate survey, and all the estimation sub-samples used in the paper.

Table A5

Value-added and Instructor Academic Rank with Inverse Probability Weighting Corrections

Dep. Variable:	*Inverse Probability Weighted VA on:*		
	Std. Course evaluation	Std. Job satisfaction	Log earnings
	(1)	(2)	(3)
Instructor academic rank (Base: Student)			
PhD	-0.017**	0.003	0.003
	(0.008)	(0.008)	(0.008)
Postdoc	0.046***	-0.014	0.007
	(0.015)	(0.012)	(0.014)
Lecturer	0.005	0.011	-0.004
	(0.008)	(0.008)	(0.010)
Assist.	0.032***	0.020**	0.013
	(0.009)	(0.008)	(0.008)
Assoc.	0.044***	0.006	-0.008
	(0.010)	(0.010)	(0.011)
Prof.	0.036***	0.011	0.014
	(0.011)	(0.009)	(0.009)
F-test inst. academic rank [p-value]	[<0.000]	[0.004]	[0.086]
R-squared	0.06	0.01	0.00
Instructors	499	478	481
Observations	1,417	1,299	1,307

This table reports WLS coefficients of regressing measures of value-added on several student outcomes on instructor academic rank, weighting by the square root of the number of students identifying each value-added estimate. Value-added measures were calculated using the inverse of the predicted response probabilities to each question as weights (Wooldridge, 2007). Predicted response probabilities were calculated from the estimates in columns 4 and 5 of Table A3 and windosorized at the 5th and 95th percentiles of their values. Heteroscedasticity-robust standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Appendix 2

Survey of Tutorial Teaching in OECD Countries

In this section, we describe the sampling procedure and show some summary statistics of the survey discussed in Section II.A. We used the Universities Worldwide Database available at https://univ.cc. to obtain a list of the population of universities in OECD countries. This database is based on the "World List of Universities 1997," which is published by the International Association of Universities and it is updated and maintained by Klaus Förster. From this database, we drew a stratified random sample without replacement from universities in OECD countries. In particular, we randomly selected three universities without replacement from each OECD country to obtain a representative picture of tutorial teaching practices in different countries. There are three exceptions to this sampling procedure. For two small countries, we could only identify contact details for fewer than three universities: two in Latvia and one in Luxembourg. Additionally, we oversampled the United States with 30 universities because they represent a 40 percent share of OECD universities. In total, our sampling population covers 4,938 universities from all OECD countries; through our survey, we contacted 139 of them. Our statistical analyses account for this complex survey design by: 1) stratifying by country, 2) including finite population corrections through stratum sampling rates, and 3) including poststratification weights constructed as the ratio of the population and the sample share of universities in the country.

We sent the survey by email to academic staff in economics, commerce, and business administration departments of the sampled universities. The email addresses were collected by a research assistant who chose academic staff who, according to their CV, are likely to speak English and have at least two years of teaching experience. To increase the response rate, we sent the survey sequentially to up to four academics per institutions. More specifically, we first sent the survey to

one academic per institution and followed up with one reminder. If the academic did not respond after the first reminder, we sent the survey to another academic in the same institution. After repeating this procedure up to four times, we got survey responses from 69 out of 139 universities, covering 31 out of 35 OECD countries.

The survey consisted of up to 18 questions and took about 5 minutes to complete. All survey questions and the survey data stripped from university identifiers is available at http://ulfzoelitz.com/research/material.

Appendix 3

Data Restrictions

Our sample period covers the academic years of 2009–10 through 2014–15. We derive our estimation sample in two steps. First, we exclude a number of observations from our estimation sample because they represent exceptions from the standard tutorial group assignment procedure at the business school. Second, we limit our estimation sample following Chetty, Friedman, and Rockoff (2014a) so that we are able to estimate instructor value-added.

Because they represent an exception to the standard tutorial group assignment procedure at the business school, we exclude the following observations:

- eight courses in which the course coordinator or other education staff actively influenced the tutorial group composition. One course coordinator, for example, requested to balance student gender across tutorial groups. The business school's scheduling department informed us about these courses.
- 21 tutorial groups that consisted mainly of students who registered late for the course. Before April 2014, the business school reserved one or two slots per tutorial group for students who registered late. In exceptional cases in which the number of late registration students substantially exceeded the number of empty spots, new tutorial groups were created that mainly consisted of late-registering students. The business school abolished the late registration policy in April 2014.
- 46 repeater tutorial groups. One course coordinator explicitly requested to assign repeater students who failed his/her courses in the previous year to special repeater tutorial groups.

- 17 tutorial groups that mainly consisted of students from a special research-based program. For some courses, students in this program were assigned together to separate tutorial groups with a more-experienced teacher.
- 95 part-time MBA students, because these students are typically scheduled for special evening classes with only part-time students

Following Chetty, Friedman, and Rockoff (2014a), and due to our own requirements for the identification of our estimates (see Section III), we exclude from our estimation sample:

- 93 tutorials with fewer than seven students, because these tutorials are considered to have too little useful variation to contribute to instructor VA estimates
- 1,410 instructor-subject observations that we did not observe for at least two periods, because we required at least two periods for each instructor-subject to construct our VA estimates
- 649 courses taught by only one instructor, because we could not identify the VA of these instructors solely using within-course variation
- 2,147 student-period observations for students who were taking more than two courses at the same time, because these students might have had to make special scheduling arrangements outside the usual system
- 71 student-year observations for which we neither observed nor could reasonably impute age