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

Jan Feld[a] Nicolás Salamanca[b] Ulf Zölitz[c]

[a] *School of Economics and Finance, Victoria University of Wellington, and IZA*
[b] *Melbourne Institute: Applied Economic & Social Research, and IZA*
[c] *University of Zurich, Department of Economics and Jacobs Center for Child and Youth Development, IZA, and Maastricht University*

May 2018

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. This is hardly 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 showing that universities can substantially reduce costs by increasingly relying on lower-ranked instructors for tutorial teaching.

Keywords: Student instructors; university; teacher value-added
JEL classification: I21, I24, J24

* Corresponding author: Ulf Zölitz. University of Zurich, Department of Economics and Jacobs Center for Child and Youth Development. Schönberggasse 1, 8001 Zurich, Switzerland. ulf.zoelitz@uzh.ch.
We thank Sophia Wagner and Dominique Gilli for providing outstanding research assistance, the Editor and three anonymous referees, Harold Cuffe, Alex 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.

1 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.[2] Tutorials – also called TA, exercise or lab sessions – are small group teaching sessions that cover material complementary to lectures.

Universities differ widely in how they staff tutorials. In some universities all tutorials are taught by students, while 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, are much more costly than student instructors, which raises the obvious question: are they 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 show how academic rank correlates with instructors' value-added (VA) on course grades, student course evaluations, and grades in follow-on courses, as well as earnings and job satisfaction after graduation. We then discuss the costs-savings potential of different staffing scenarios that rely on increasing the share of lower-rank instructors.

[1] See, for example, Ehrenberg (2012) about the increase of adjunct professors in the U.S. 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 small positive effect on students' subsequent course enrolment.

[2] To learn more about the prevelance of tutorial teaching, we conducted a small survey among OECD universities. The survey results suggest that around 47 percent of OECD universities use tutorials, and in these universities tutorials make up around 28 percent of students contact hours. See Section 2.1 and Appendix B1 for a detailed description of the survey methodology and results. The survey data is available online at http://ulfzoelitz.com/research/material.

Our results show that instructors' academic rank is, overall, unrelated to students' academic outcomes. The most effective instructors – postdocs – add less than one percent of a standard deviation more to students' grades than student instructors. All other types of instructors, including full professors, are not significantly more effective than students in tutorial teaching. Importantly, these findings are not driven by a lack of statistical power. We can rule out differences between instructor types as small as one percent of a standard deviation of a grade. Instructors' academic rank is also unrelated to students' grades in follow-on courses, where our results are also precisely estimated.

Looking at non-academic outcomes, we find that instructors with higher academic ranks add more value to students' course evaluations. These differences, however, are also small. Students taught by a full professor, for example, evaluate the course only 3 percent of a standard deviation more positively than students taught by a student instructor. Finally, using matched survey data on university graduates, we find some evidence that PhD students and professors have a small positive effect on students job satisfaction after graduation. We find no systematic relationship between academic rank and students' earnings. 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 and subsequent earnings, and only small negative effects on course evaluations and job satisfaction. 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 which shows the savings potential under different tutorial staffing scenarios. In the most extreme scenario where all tutorials are taught by student instructors, wage costs for the average tutorial can be reduced by 47 percent for the bachelor's and by 55 percent for the master's program. Under a more conservative scenario

where 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 31 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. The estimated relationship between academic rank and instructor effectiveness, however, 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 (2016) find that instructors' effectiveness in one algebra course at the University of Phoenix is unrelated to their salary.

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 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 compared to native graduate TAs, and reach opposing conclusions. Borjas (2000) finds that

[3] Another extensive 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 (e.g. Chetty, Friedman, and Rockoff, 2014). There are, however, 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 the first courses in that subject was taught by a PhD student.

foreign-born TAs negatively affect student grades, whereas Fleisher et al. (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 compares the effectiveness of tutorial instructors with different academic ranks.

We make three main contributions. First, we are the first study that focuses on instructors' effectiveness in tutorial teaching. Since tutorials are a critical part of most students' university education, our study fills an important knowledge gap in the literature on teacher effectiveness. Second, our rich dataset 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 long-run. Third, we discuss potential savings under different staffing scenarios. The size of these potential savings, together with our main results, will help university administrators make better informed staffing decisions.

2 Background and Data

2.1 Tutorial Teaching in OECD Countries

In order to understand how common tutorial teaching is and how it differs between institutions, we conducted a small email survey among universities of 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 B1.

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

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
Students scheduled per tutorial	49	22.23	0	140
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 is teaching tutorials:				
Only students	49	0.25	0	1
A mix of both	49	0.46	0	1
Only professors	49	0.29	0	1
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
What happens in tutorials in undergraduate courses:				
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
What happens in tutorials in graduate courses:				
Instructor stands in front of the class and explain	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 Section B1 in the Appendix.

indicate that over 62 percent of universities in OECD countries offer tutorials at undergraduate or graduate level. The average tutorial group size is 23 students. In universities where tutorials are used, students spend about 28 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 29 percent of all universities use only student or PhD student instructors, whereas 49 percent use a mixture of student and higher-ranked instructors, such as assistant, associate and full professors. About 22 percent of universities staff their tutorials exclusively with professors.

The fact that so many institutions use higher-ranked instructors begs the question of whether these more expensive instructors are also more effective.

2.2 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 2009/2010 to 2014/2015.[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 (e.g., Principles of Microeconomics) and 'course' to refer to a subject-year-period combination (e.g., 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, 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.

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

Tutorials are organized in groups of up to 16 students who are assigned to one instructor. In these tutorials, students discuss assigned readings or go over 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 instructors is to guide tutorial meetings 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 group composition we observe in our data therefore closely reflects the actual tutorial group compositions.

Our institutional background and the method we use to calculate instructor value-added (VA) impose some sample restrictions. Since we aim to distinguish course effects from instructor effects, we limit our sample to courses which were taught by at least two different instructors. Moreover, we follow Chetty et al. (2014) for our VA calculation (see Section 3) in three ways. First, we limit our estimation sample to instructors teaching the same subject in at least two periods since 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 they teach different subjects, we consider each instructor-subject combination

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

as separate instructor. Thus, the same person teaching Microeconomics and Macroeconomics is counted as two separate instructors in our data. We discuss these, and a few other sample restrictions in greater detail in Appendix B2. Our final estimation sample consists of 559 instructor-subject observations (which we refer to as instructors from now on), 651 different courses, and 12,257 student-course observations.[7]

2.3 Tutorial Instructors

Students in the same course can be assigned 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. There is also substantial variation by academic rank in the number of courses for which instructors taught tutorials in our data, ranging from an average of 2.27 courses taught by student instructors to 3.68 taught by 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

[7] Table A1 in the Appendix shows how the sample courses differ from the non-sample 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 estimation of instructor VA (see Section 3). However, they are important for interpreting our results, since they define the population of courses which we use to estimate our results.

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
Courses taught	2.27	2.70	3.23	3.54	3.56	3.23	3.68
Tutorials taught	7.13	7.30	9.17	10.71	8.76	7.77	8.97
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 meeting	€ 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 is based on our estimation sample, comprising 48,842 observations from 12,257 students who took 651 different courses, taught by 559 instructors over 24 teaching periods between 2009 and 2014. Total wage costs per tutorial meeting include paid preparation time.

academic rank. PhD students are more likely to teach mathematical courses, whereas postdocs and lecturers teach more first-year courses.

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), 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 has to pay €134 Euros more for a tutorial taught by a full professor than for a tutorial taught by a student. These wage cost differences are themselves a lower bound for the total cost differences between instructor rank, since they ignore overhead and hiring costs,

which are likely also larger for higher-ranked instructors. They are also not dampened by in-kind benefits received by lower-ranked instructors since student instructors do not receive tuition waivers, or any other non-monetary benefits besides their salary.

Besides teaching, instructors also differ in terms of their contractual terms and their non-teaching responsibilities. Professors, lecturers, and postdocs work on permanent or temporary contracts. While lecturers mainly teach, professors and postdocs also do academic research and fulfill administrative duties. PhD students and student instructors pursue their own studies parallel to teaching. PhD students are typically required to teach 20 percent of their time. Student instructors are often hired when there are not enough regular staff or PhD students available to cover a course's teaching load, and therefore they disproportionally teach large bachelor's courses. They are typically hired by an administration staff member mainly based on their grades, previous experience with the course at hand, and a sufficient command of English, the language of instruction for all courses. Their contracts are always part-time and their only teaching obligation is tutorial teaching.

2.4 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 earnings and job satisfaction after graduation. Figure 1 shows the distributions of all these variables and reports their means and standard deviations in our estimation sample.

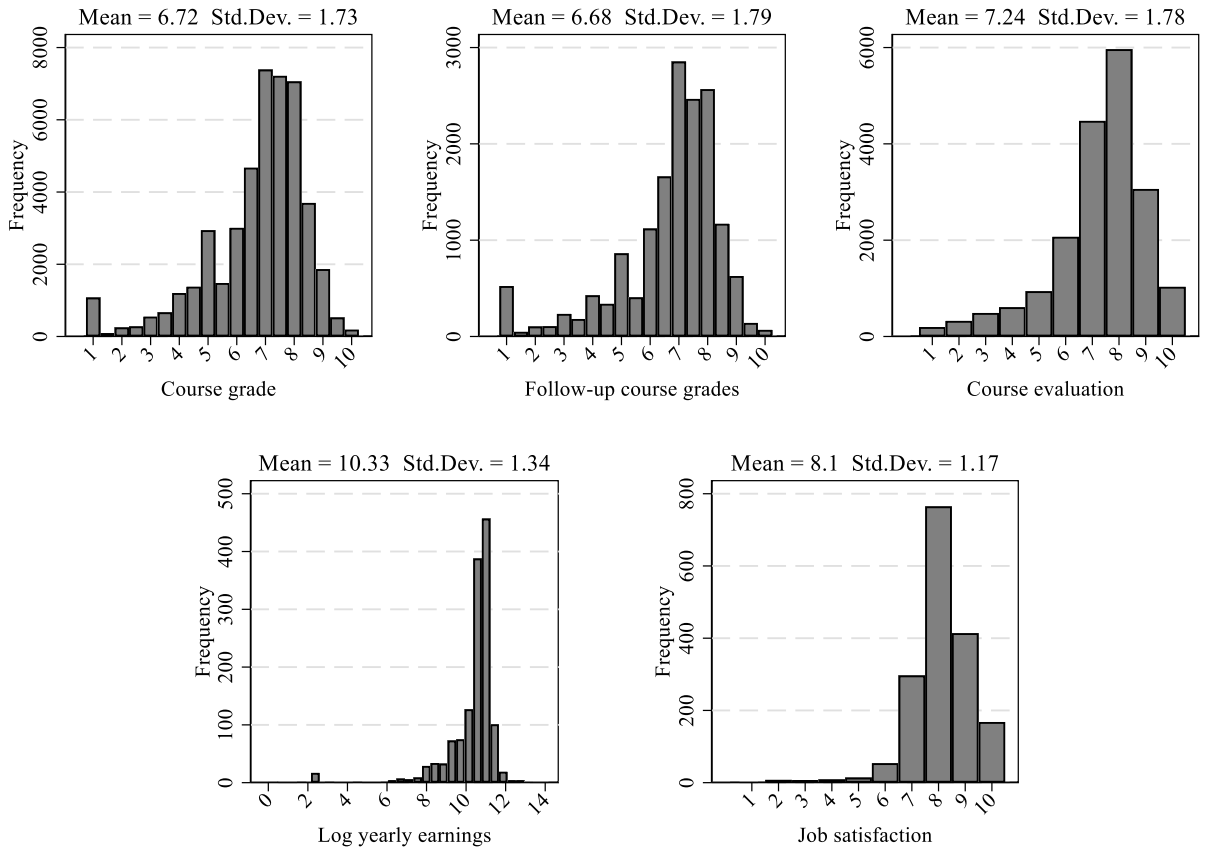

Figure 1
Distribution of Student Outcomes

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. Exams are usually marked by all instructors involved in the course, 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 to five, and the second instructor may grade questions six to 10 for all students. Much of the exam is therefore not graded by the students' own instructor.[8] The other graded components and their weights vary across

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

courses, and some of these components, such as group work or tutorial participation, are directly graded by the students' instructor.

In our data we only observe final course grades and one potential concern is that there are differences in grading standards by academic rank. We argue in Section 4 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 the Appendix).

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 towards 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 38 percent, and there is some evidence of selective non-response related to instructor rank. The fourth column of Table A3 in the Appendix shows that PhD students and full professors achieve significantly lower survey response rates than student instructors. We show in Section 4 that none of our results change once we account for this selectivity.

To collect data on students' earnings and job satisfaction post-graduation, we surveyed students who obtained their undergraduate degree between 2011 and 2016. In this survey, we measure earnings as the pre-tax yearly earnings from their main occupation, and job satisfaction on a 10-point scale, with 10 being most satisfied. The survey response rate is 37 percent, and in the matched data we have information on earnings and job satisfaction from 1,737 students in our estimation sample. There is weak evidence of some selective non-response related to instructor

rank in this survey. The last column of Table A3 shows that students taught by PhD students have a somewhat higher response rate, yet instructor ranks are not jointly significant in predicting survey response. We show in Section 4 that correcting for this selectivity does not change our results.

2.5 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.[9] For all bachelor students, this assignment was unconditionally random until the academic year of 2009/2010. From 2010/2011 onwards, 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 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 & Zölitz, 2017; Zölitz & Feld, 2017; Mengel et al., forthcoming).

[9] Courses are usually scheduled in a way that avoid 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. To account for potentially non-random assignment due to other courses taken at the same time, we control for fixed effects for all combinations of courses that students take in each period. 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 fall 2015, students could opt-out of participating in tutorials which start at 6:30 pm. 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.

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

Table 3 shows that academic rank is not systematically related to any of these five pre-treatment 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 only see that postdocs tend to teach younger students. This difference is, however, small and likely due to chance given that we calculate 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 instructor's academic rank is not systematically related to pre-treatment student and tutorial group characteristics.

Table 3
Balancing of Pre-Determined 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	0.205
	(0.043)	(0.017)	(0.053)	(133.151)	(0.127)
Postdoc	0.036	-0.005	-0.138**	-25.050	0.191
	(0.061)	(0.025)	(0.069)	(211.064)	(0.156)
Lecturer	0.029	-0.007	-0.004	-108.071	0.146
	(0.037)	(0.015)	(0.044)	(132.700)	(0.110)
Assist.	0.045	-0.014	-0.083	-180.238	0.092
	(0.043)	(0.018)	(0.054)	(151.702)	(0.127)
Assoc.	0.060	-0.026	0.032	-177.155	0.223
	(0.061)	(0.026)	(0.078)	(186.527)	(0.175)
Prof.	0.043	-0.011	-0.001	-146.954	0.069
	(0.066)	(0.028)	(0.091)	(225.900)	(0.225)
Instructor gender, nationality, experience:	✓	✓	✓	✓	✓
Tutorial schedule FE:	✓	✓	✓	✓	✓
Course FE:	✓	✓	✓	✓	✓
F-test all inst. characteristics [p-value]	[0.927]	[0.871]	[0.129]	[0.977]	[0.785]
F-test inst. academic rank [p-value]	[0.961]	[0.874]	[0.066]	[0.922]	[0.685]
R-squared	0.70	0.05	0.49	0.01	0.75
Instructors	1,486	1,490	1,490	1,490	1,481
Observations	44,616	48,842	48,842	48,842	3,782

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

3 Estimation of Instructor Value-added

3.1 Empirical Strategy

We estimate the effect of academic rank on instructor effectiveness by testing whether instructors' academic ranks predictes 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 et al. (2014) 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 students' grades 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.

The third step is the true innovation of the Chetty et al. method. It brings three improvements over 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 if grade VA predicts current grades without worrying about spurious results. Second, using all other years' averaged residuals as *separate* predictors allows residuals closer apart in time to be better predictors than residuals that are further apart. Chetty et al. show that this "drift adjustment" improves out-of-sample VA forecasts. Third, using predicted VA instead of residual averages shrinks VA estimates towards their mean. Since 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 towards 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 their age and dummies for their gender, nationality, whether they are bachelor's students, exchange students, whether a student is repeating the course, and students 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 students registered late for the course.[10] 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:

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

In this error structure, time-varying course-specific unobserved heterogeneity, δ_{ct}, is allowed to 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

[10] 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, where we cannot observe it, and interact a dummy for this period with our GPA measure. All these choices follow Chetty et al. (2014, p.13).

instructors systematically sorting into courses. And, 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 $\widetilde{grade}_{icjt} = grade_{icjt} - \overline{grade}_{jt}$ on $\tilde{X}_{icjt} = X_{icjt} - \bar{X}_{jt}$, where $\overline{grade}_{jt}$ and $\bar{X}_{jt}$ 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. 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 $\tilde{X}_{icjt}$, since we would otherwise attribute part of the instructor effect to these covariates (Chetty et al., 2014, p.6). By adding instructor fixed effects, thus, 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 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 occurs (see Section 2.2). 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.

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

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

These residuals are the basis for the grade VA estimates. We aggregate the residuals to instructor-time weighted averages, $\overline{grade}^{*}_{jt}$ using Chetty et al.'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:

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

where $\hat{\psi}_k$ is the bivariate OLS coefficient from regressing $\overline{grade}^{*}_{jt}$ on $\overline{grade}^{*}_{jk}$ for $k \neq t$. However, we restrict our residual autocovariances to be constant after $k > 4$, which is similar to imposing some restrictions on the OLS coefficients used for constructing predictions. This type of covariance restriction is illustrated clearly in Chetty et al. (2014, p.2608). 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, $\hat{\psi}_k$, used for creating our final VA measures.[11]

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

[11] Since 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 4-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 since we use within-course transformation in the construction of the residuals (step 1).

$$VA_{jt}^{y} = \gamma^{y} VA_{jt}^{grade} + \epsilon_{jt}^{y}, \quad (5)$$

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

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

$$VA_{jt}^{y} = \theta^{y} Rank_{jt} + e_{jt}^{y}, \quad (6)$$

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 students 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 all these regressions, we cluster our standard errors at the instructor level to account for potential correlation of the error term within instructors.

3.2 Estimates of Instructor Value-added

Table 4 summarizes our short and longer-run 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 0.04, which means that an instructor who is one standard deviation more effective increases students' grades by 0.04 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 if 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.

Table 4
Summary Statistics of Value-added Estimates

	Obs	No. inst.	Std. Dev.	*Percentile:* Min.	1st	50th	99th	Max.	F-Test of Instructor FE
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Value Added estimates:									
Course grade	1,432	502	0.041	-0.148	-0.113	0.001	0.112	0.214	[0.000]
Follow-on grade	958	346	0.057	-0.339	-0.165	0.003	0.132	0.236	[0.000]
Course evaluation	1,417	499	0.121	-0.666	-0.282	0.004	0.325	0.418	[0.000]
Log earnings	1,304	481	0.052	-0.634	-0.163	0.000	0.132	0.612	[0.991]
Job satisfaction	1,317	482	0.066	-0.842	-0.163	0.000	0.186	0.807	[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 VA 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. F-Test based on regular standard errors.

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 et al., 2016; Carrell & West, 2010; Hoffmann & 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 & Rivkin, 2010).

Our estimated standard deviation in follow-on grade VA is 0.057 grade points, which is quite similar in size to our grade VA estimates. The standard deviation of VA on course evaluations, which are like grades measured on a 10-point scale, is about three times as large with 0.121 points. For both measures, we also observe significant heterogeneity between instructors. The standard deviation of job satisfaction VA is at 0.066 points on a 10-point scale more similar to the standard deviation of course evaluation VA. The standard deviation in log-earnings VA is quite large at 0.052, suggesting that a one standard deviation more effective instructor adds 5 percent to students' earnings. However, for log-earnings VA and job satisfaction VA we cannot reject the null hypothesis that individual instructors do not affect these labor market outcomes.

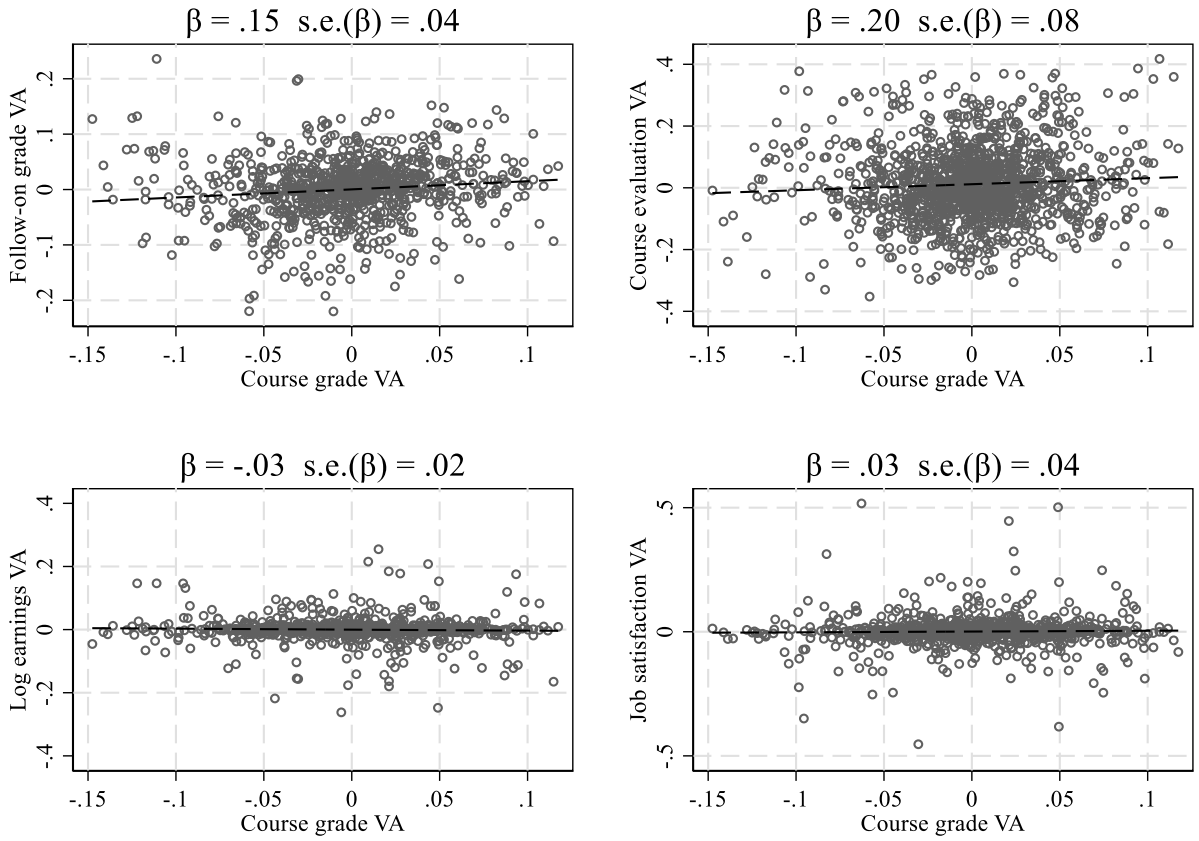

Figure 2
The Relation Between Different Instructor Value-added Estimates

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 evaluation, or their earnings and job satisfaction in the labor market? Figure 2 answers this question by showing scatterplots of the relation between grade VA and all other VA estimates, as well as the bivariate regression coefficients underlying their linear relations. Instructors' effectiveness in raising students' current and future grades are significantly related. An instructor who adds one point to students' current grades also adds 0.15 points to their follow-on grades. This persistence of grade VA contrasts 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 a small degree of achievement VA persistence in primary and secondary education. There is also a relationship between grade VA and course evaluation VA; instructors who add one more point to their student's grades get 0.20 points better course ratings. None of the relationships between the other VA measures shown in Figure 2 are statistically or economically significant, suggesting that instructor VA on grades does not persit onto students' labor market outcomes after graduation.

4 Academic Rank and Instructor Value-added

4.1 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 onwards 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 instructor's 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 or Lecturers and postdocs, and it amounts to little more than one percent of a standard deviation in grades. These small differences, though, are precisely estimated. We can, for example, rule out that

Table 5
Value-added and Instructor Academic Rank

	Value Added on:				
Dep. Variable:	Std. Course grade	Std. Follow-on grade	Std. Course evaluation	Log earnings	Std. Job satisfaction
	(1)	(2)	(3)	(4)	(5)
Instructor academic rank (Base: Student)					
PhD	-0.003	-0.008**	-0.013*	-0.004	0.012**
	(0.002)	(0.004)	(0.006)	(0.004)	(0.005)
Postdoc	0.009***	0.007	0.034***	-0.001	0.006
	(0.003)	(0.006)	(0.013)	(0.006)	(0.008)
Lecturer	-0.003	-0.007	0.004	-0.005	0.009
	(0.002)	(0.004)	(0.007)	(0.004)	(0.006)
Assist.	0.004*	0.001	0.024***	-0.001	0.011**
	(0.003)	(0.004)	(0.007)	(0.004)	(0.006)
Assoc.	0.004	0.003	0.032***	0.019	0.016**
	(0.003)	(0.005)	(0.008)	(0.013)	(0.007)
Prof.	-0.002	-0.005	0.032***	-0.000	0.013**
	(0.003)	(0.006)	(0.008)	(0.004)	(0.006)
F-test inst. academic rank [p-value]	[0.000]	[0.001]	[0.000]	[0.464]	[0.254]
R-squared	0.02	0.02	0.05	0.01	0.00
Instructors	502	346	499	481	482
Observations	1,432	958	1,417	1,304	1,317

*This table reports OLS coefficients of regressing measures of value added on several student outcomes on instructor academic rank. Heteroscedasticity-robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1*

the difference in effectiveness between full professors and student instructors is more than one percent of a standard deviation in grades.

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. The estimated effect sizes, however, are again small. The largest coefficients suggest that postdocs, associate professors and full professors add 3.4 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 response 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 non-responding students would have been for each instructor rank. To correct for potential bias due to selective response, we calculate course evaluation VA giving more weight to observations who have a lower predicted probability of responding to the evaluation survey following Wooldridge (2007). 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 dependent variable.[12]

Finally, we estimate instructor VA on earnings – measured in pre-tax log-points – and job satisfaction, both of which are important labor market outcomes. We thus test the possibility that instructors affect students' labor market outcomes by, for example, giving career advice, even if

[12] The first column of Table A4 in the Appendix shows the main results with inverse probability weighted (IPW) VA estimates. 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 percentile. Under the assumption that our rich set of observed characteristics can inform the selection process (i.e., "coarsened at random" selection), the IPW estimator is consistent and more efficient than OLS (Wooldridge, 2007).

the do not affect their grades and course evaluations. For log-earnings VA we find no significant differences between instructors of different academic ranks. Our log-earnings VA estimates are also somewhat noisier than those on university outcomes. This is particularly true for the effect of assistant professors, where we would not be able to detect a wage difference of 3.8 percent versus student instructors based on the coefficient's standard error. However, the coefficients for all the other instructor ranks are precise enough that we can rule out earnings differences of around one percent compared with student instructors. We therefore view our results as relatively powerful evidence that instructor rank does not affect future earnings.

Instructor rank is also not jointly predictive of job satisfaction VA, with an insignificant F-test on instructor rank, and the largest difference between instructors (associate professors versus student instructors) being a mere 1.6 percent of a standard deviation. By inspecting the coefficients, however, it seems that PhD students and professors of all levels have a higher job satisfaction VA than student instructors, postdocs and lecturers. Recall professors rank also tend to add more to students' course evaluations. These results together are consistent with the idea that higher-ranked instructors increase the non-pecuniary benefits of education for students. However, we are hesitant to draw strong conclusions on the matter since our evidence is statistically too weak.[13]

Taken together, 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 earnings. Our results on course evaluation and job satisfaction are suggestive of some non-pecuniary benefits of higher-ranked instructors, yet

[13] Our estimates for log-earnings VA and job satisfaction 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 column of Table A4 in the Appendix.

the magnitudes of these effects are minute. These results are consistent with the existing literature which has repeatedly shown that observable instructor characteristics are not strong determinants of differences in teacher VA (see Koedel, Mihaly and Rockoff (2015) for a recent review).

4.2 Heterogeneity by Subject Type

While we do not find meaningful differences in VA by academic rank, these average effects may still hide important heterogeneity by subject type. In this sub-section we test if higher ranked instructors matter more for mathematical and non-first-year subjects, since 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 provide 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 courses 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 graded components, like participation grades or presentation grades, which are graded by the tutorial instructor. If grading biases are driving our results, we would therefore expect larger academic rank differences in course 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.

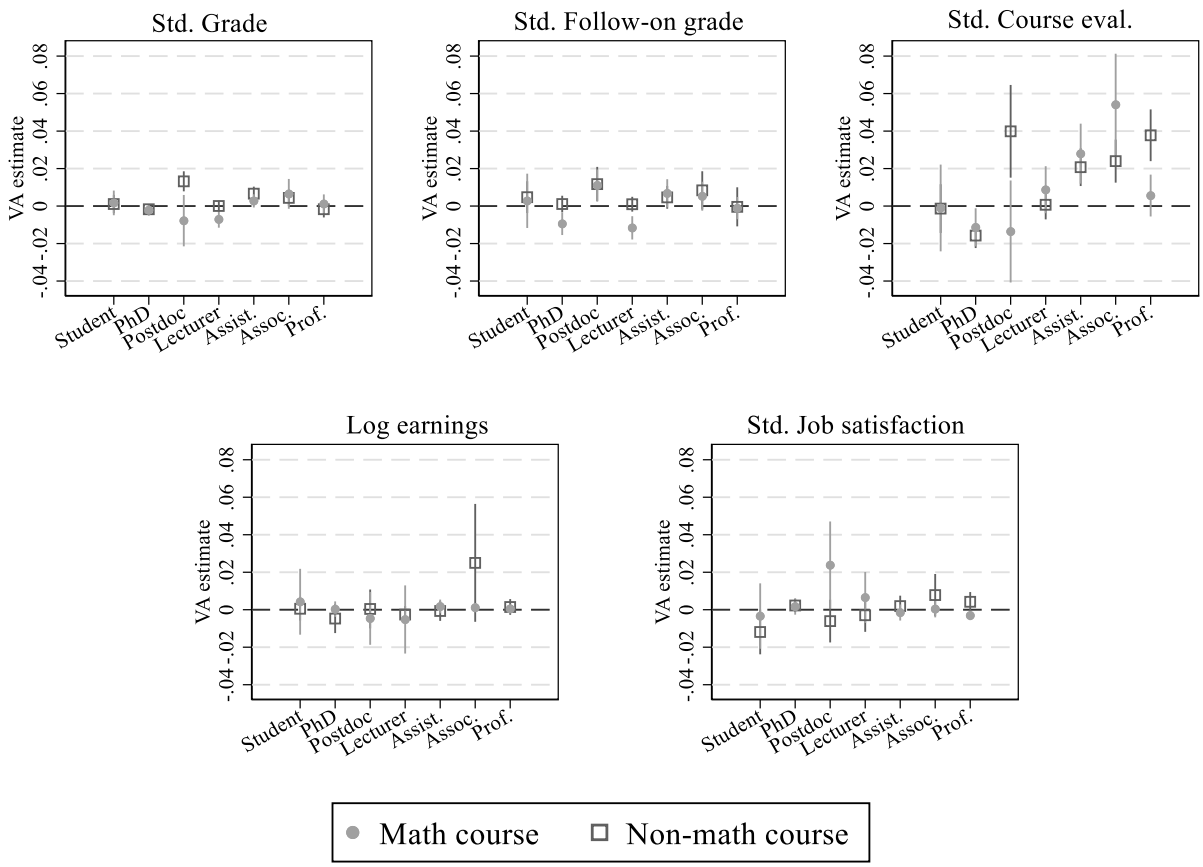

Figure 3
Value-added and Instructor Academic Rank in Mathematical and Non-Mathematical Subjects

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 one 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. For log-earnings VA, the only economically significant estimate suggests that associate professors increase students'

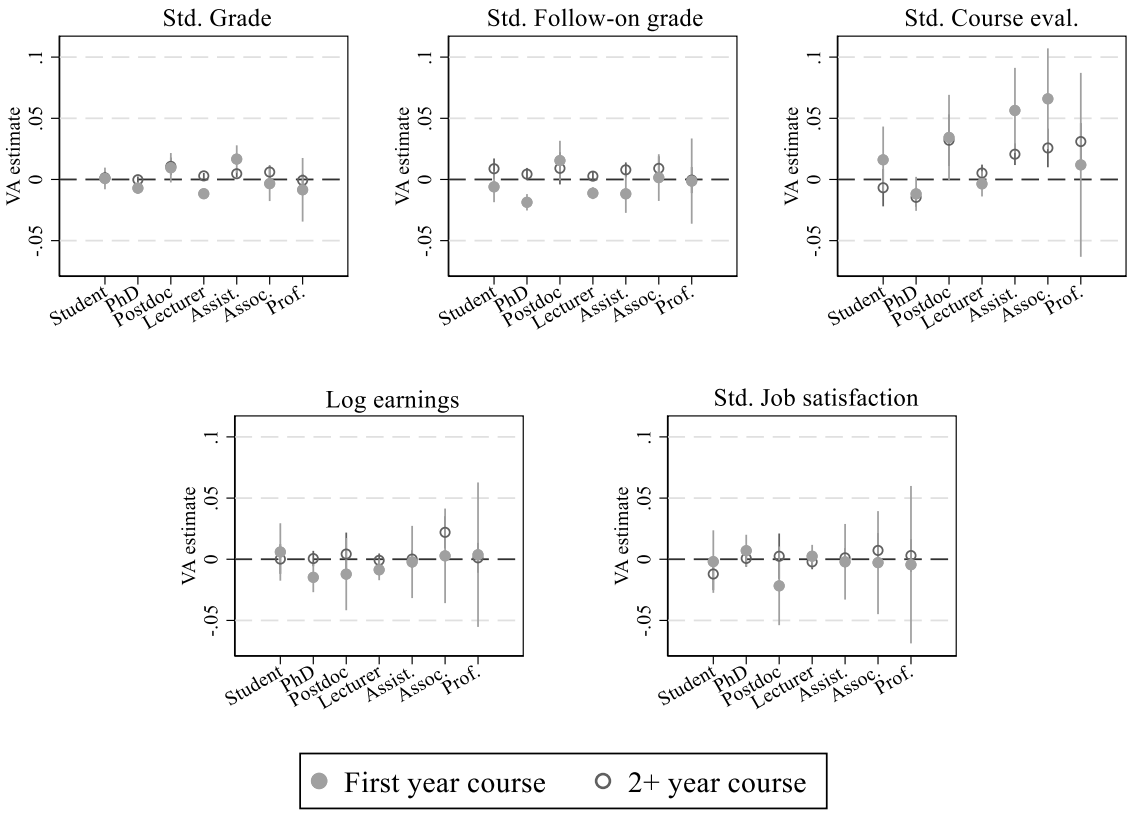

Figure 4
Value-added and Instructor Academic Rank in First-Year and Non-First-Year Subjects

earnings in non-math subjects, but neither this nor any other estimate is statistically significant. We also see no systematic or economically meaningful heterogeneity for job satisfaction VA. Overall, we find little evidence that the effect of academic rank on student outcomes differs by math course content.

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. While 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 log-earnings VA, though in this sub-group analysis some of the log-earnings VA estimates become imprecise. Taken together our results suggest that there exists little meaningful heterogeneity between first-year and non-first-year subjects.

5 Saving 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. Bulding on these results, we explore in this section the potential gains of using lower-ranked instructors. We do this by conducting an accounting exercise which showcases potential savings from changing the staff composition. The 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 49 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 arrangement and wage costs. Nevertheless, this exercise emphasizes the potential magnitude of two different cost-saving policies that can be implemented in many universities around the globe.

Table 6 shows the proportion of tutorials that are taught by instructor of different academic rank, and the average wage cost per tutorial in the status-quo, and in two alternative scenarios where we replace higher-ranked instructors with lower-ranked ones. In the status-quo, lecturers and professors of any rank teach 50 percent of all bachelor's and 74 percent of all master's tutorials. The wage cost per tutorial is thus €105 at the bachelor-level, and €125 at the master-level.

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	21%	8%	100%	100%	63%	8%
PhD	€ 79	25%	15%	0%	0%	25%	15%
Postdoc	€ 93	4%	3%	0%	0%	0%	3%
Lecturer	€ 126	25%	16%	0%	0%	5%	16%
Assist.	€ 126	13%	28%	0%	0%	4%	28%
Assoc.	€ 173	8%	15%	0%	0%	2%	15%
Prof.	€ 190	4%	13%	0%	0%	1%	13%
Average wage costs per tutorial		€ 105	€ 125	€ 56	€ 56	€ 72	€ 125
Total savings potential				47%	55%	31%	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 since the actual cost reduction for substituting senior instructors is underestimated.

In our first alternative scenario, we calculate costs for the case where 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 all tutorials are taught by students. The average wage costs per tutorial in this scenario decreases to €56 for both bachelor's and master's tutorials. This is a 46 percent decrease in the wage costs for the average bachelor's tutorial, and a 55 percent decrease for the average master's tutorial.

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 than 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 effect on later-life earnings. However, since each instructor affects around 15 students per period, even small earning penalties on student instructors would quickly accrue to large costs born by students later on. We are thus explicitly interpreting the evidence in Table 5, which shows that we can reject the existence of quite small earning effects, as evidence of no differences in earnings between instructors of different ranks, and lean on this interpretation for the current cost exercise. Finally, we are not monetizing the small effect on course evaluations of replacing higher-ranked instructors by student instructors. Course evaluations could, however, have a monetary value to the business school, and moreover a drop in course evaluations could be reflecting a loss of non-pecuniary value of education for the students. Yet, since we have no prior that allows us to include either of these features of course evaluation as a monetary cost, we refrain from considering them in our analyses.

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 since our estimates shown in Section 4 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, & Taylor, 2016). Third, our VA estimates are mainly driven by bachelor's courses and our results may not generalize to all master's courses. Many of the smaller master's courses which are excluded from our estimation sample because they only had one instructor, for example, may be too technical for student instructors.

In our second alternative scenario, we keep these caveats in mind and simulate a counterfactual staff assignment where we do not change the staff composition in master's courses, keep the status-quo share of PhD students in bachelor's course, and allow the highest-ranked instructors in each bachelor's course to teach one tutorial. In this scenario, the average wage costs decrease from €105 to €72 for bachelor-level tutorials – a 31 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. In all cases, however, universities should still consider the opportunity costs of time for higher-ranked instructors. These opportunity costs likely differ between instructors. There might be, for example, some research inactive and tenured professors whose most valuable use of time is teaching tutorials. We generally believe, however, that most professors are more valuable doing other activities like research. While 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 which seek to reduce costs or want to give their professors more time for research.

6 Conclusion

Universities around the world have very different policies in 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, their grades in follow-on courses, the evaluations students give to the courses, and students' subsequent earnings and job satisfaction. We show that, despite substantial differences in formal

qualification 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 difference in teaching effectiveness between full professors and students as large as one percent of a grade standard deviation. We find evidence that professors receive marginally higher course evaluations, and lead students to jobs where they are more satisfied. Yet, these estimates are economically miniscule. We find no evidence that academic rank is systematically related to students' earnings, although the estimate for associate professors is less precise.

There might be, of course, differences in teaching effectiveness that we do not capture with our broad range of outcomes. Professors, for example, might be better at dealing with students' family problems and mental health issues. Lower-ranked instructors, however, could also offer benefits to their students which are unobserved by us. 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. However, 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. For instance, 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 instructors' main role is explaining the course material. This is an important empirical question, and we believe more research is needed to see if our findings replicate in other settings.

Taken together, however, our results raise an important question: Is tutorial teaching really the best use of a professor's time? Our findings suggest that instead of asking professors to teach tutorials, universities should consider giving them additional time for more productive activities like research.

References

Bettinger, E. P., & Long, B. T. (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

Bettinger, E. P., Long, B. T., & Taylor, E. S. (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

Borjas, G. J. (2000). Foreing-born teaching assistants and the academic performance of undergraduates. *American Economic Review*, *90*(2), 355–359.

Braga, M., Paccagnella, M., & Pellizzari, M. (2016). The impact of college teaching on students' academic and labor market outcomes. *Journal of Labor Economics, 34*(3), 781-822.

Carrell, S. E., & West, J. E. (2010). Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy*, *118*(3), 409–432. https://doi.org/10.1086/653808

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). 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

De Vlieger, P., Jacob, B., & Stange, K. (2016). Measuring instructor effectiveness in higher education, *NBER Working Paper* No. 22998. https://doi.org/10.3386/w22998

Ehrenberg, R. 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, J., Salamanca, N., & Hamermesh, D. S. (2016). Endophilia or exophobia: Beyond discrimination. *Economic Journal*, *126*(594). https://doi.org/10.1111/ecoj.12289

Feld, J., & Zölitz, U. (2017). Understanding peer effects: On the nature, estimation, and channels of peer effects. *Journal of Labor Economics*, *35*(2). https://doi.org/10.1086/689472

Figlio, D. N., Schapiro, M. O., & Soter, K. B. (2015). Are tenure track professors better teachers? *Review of Economics and Statistics*, *97*(4), 715–724. https://doi.org/10.1162/REST_a_00529

Fleisher, B., Hashimoto, M., & Weinberg, B. A. (2002). Foreign GTAs can be effective teachers of economics. *Journal of Economic Education*, *33*(4), 299–325. https://doi.org/10.1080/00220480209595329

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review*, *100*(2), 267–271. https://doi.org/10.1257/aer.100.2.267

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of public economics*, *95*(7-8), 798-812.

Hoffmann, F., & Oreopoulos, P. (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, B. A., Lefgren, L., & Sims, D. P. (2010). The persistence of teacher-induced learning. *Journal of Human Resources*, *45*(4), 915–943. https://doi.org/10.1353/jhr.2010.0029

Kane, T. J., & Staiger, D. O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. *NBER Working Paper* No. 14607. https://doi.org/10.3386/w14607

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, *47*, 180–195. https://doi.org/10.1016/j.econedurev.2015.01.006

Lusher, L., Campbell, D., & Carrell, S. (2018). TAs like me: Racial interactions between graduate teaching assistants and undergraduates. *Journal of Public Economics*, *159*(2), 203-224. https://doi.org/10.1016/j.jpubeco.2018.02.005

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, U., & Feld, J. (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 A1
Additional Tables

Table A1
Comparison of Sample vs Non-Sample 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	7.10	-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	0.10	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	10.77	2.65	8.12
No. students per tutorial	12.64	11.44	1.20

This table is based on data from 111,481 observations from 14,051 students who took 1,279 courses, taught by 2,054 instructors over 24 teaching periods between 2009 and 2014.

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)	160	32	40	160	80	80	80
FTE Standard teaching load	*flexible*	0.20	0.25	1	0.50	0.50	0.50
Hourly wage	€ 14.1	€ 19.9	€ 23.2	€ 31.5	€ 31.5	€ 43.2	€ 47.5
Hours per tutorial meeting 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 meeting	€ 56.28	€ 79.48	€ 92.85	€ 125.85	€ 125.85	€ 172.70	€ 189.98

Note: Monthly gross wages are assumed to be in the lowest pay scale of the instructor type which provides with 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*	0.019
	(0.009)	(0.009)	(0.022)	(0.019)	(0.018)
Postdoc	-0.004	-0.005	0.019	-0.041	-0.010
	(0.011)	(0.012)	(0.036)	(0.034)	(0.026)
Lecturer	-0.003	0.001	0.002	-0.006	0.007
	(0.008)	(0.007)	(0.021)	(0.017)	(0.015)
Assist.	-0.000	0.003	0.003	-0.016	-0.007
	(0.009)	(0.008)	(0.026)	(0.020)	(0.017)
Assoc.	-0.002	0.004	0.007	-0.025	0.013
	(0.011)	(0.010)	(0.038)	(0.024)	(0.021)
Prof.	-0.002	0.001	0.002	-0.061**	-0.011
	(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]	[0.346]
R-squared	0.07	0.67	0.08	0.09	0.14
Instructors	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. Non-parametric bootstrap standard errors clustered at the instructor-by-time level in parentheses. *** p<0.01, ** p<0.05, * p<0.1*

Table A4

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

	Inverse Probability Weighted VA on:		
Dep. Variable:	Std. Course evaluation	Log earnings	Std. Job satisfaction
	(1)	(2)	(3)
Instructor academic rank (Base: Student)			
PhD	-0.022*	-0.003	0.015**
	(0.011)	(0.004)	(0.006)
Postdoc	0.059***	-0.003	0.008
	(0.022)	(0.006)	(0.009)
Lecturer	0.008	-0.006	0.014**
	(0.012)	(0.004)	(0.007)
Assist.	0.042***	-0.003	0.014**
	(0.013)	(0.004)	(0.007)
Assoc.	0.055***	0.016	0.020**
	(0.014)	(0.012)	(0.008)
Prof.	0.055***	-0.001	0.016**
	(0.014)	(0.004)	(0.007)
F-test inst. academic rank [p-value]	[0.000]	[0.482]	[0.236]
R-squared	0.05	0.01	0.01
Instructors	499	481	482
Observations	1,416	1,304	1,314

*This table reports OLS coefficients of regressing measures of value added on several student outcomes on instructor academic rank dummies. VA 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 3 and 4 of Table A3 and windosorized at the 5th and 95th percentiles of their values. Heteroscedasticity-robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1*

Appendix B1

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 2.1. We used the Universities Worldwide Database available at https://univ.cc/. to get 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 databased 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 from this sampling procedure. For two small countries we could only identify contact details for fewer than three universities: two in Latvia, one in Luxembourg. Additionally, we oversampled the US with 30 universities, since they represent with 40 percent a large share of OECD universities. In total, our sampling population covers 4,938 universities from all OECD countries, and through our survey we contacted 139 of them. Our statistical analyses account for this complex survey design by i) stratifying by country, ii) including finite population corrections through stratum sampling rates, and iii) including post-stratification 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 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 consists 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 B2
Data Restrictions

Our sample period comprises the academic years 2009/10 until 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 et al. (2014) 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.
- 268 evening tutorials comprised of students who did not opt out of evening education.
- 21 tutorial groups that consisted mainly of students who registered late to the course. Before April 2014, the business school reserved one or two slots per tutorial group for students that

registered late. In exceptional cases where the number of late registration students substantially exceeded the number of empty spots, new tutorial groups were created that mainly consist 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 courses in the previous year to special repeater tutorial groups.
- 17 tutorial groups that consist mainly of students from a special research-based program. For some courses, students in this program are assigned together to separate tutorial groups with more experienced teacher.
- 95 part-time MBA students, since these students are typically scheduled for special evening classes with only part-time students.

Following Chetty et al. (2014), and due to our own requirements for the identification of our estimates (see Section 3) we exclude from our estimation sample:

- 93 tutorials with fewer than seven students, since these tutorials are considered to have too little useful variation to contribute to instructor VA estimates.
- 1,410 instructor-subject observations that we do not observe for at least two periods, since we require at least two periods for each instructor-subject to construct our VA estimates.
- 649 courses taught by only one instructor, since we cannot 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, since these students might have had to make special scheduling arrangements outside the usual system.
- 71 student-year observations for which we neither obseve nor can reasonably impute age.