Foreign instructors and student STEM outcomes

Daniela Morar *

November 11, 2019

Preliminary draft- do not cite

Abstract

The past decades have experienced an increase in the enrollment of foreign-born students in U.S. STEM (Science, Technology, Engineering, and Math) graduate programs. This paper investigates whether having a foreign teaching assistant (TA) in a STEM class affects the outcomes of U.S. undergraduate students. I consider both subjective outcomes (the median evaluation scores) and objective ones (the students' course outcomes). I use administrative data from a large public university where TAs are conditionally-randomly allocated to classes. I find that TAs from countries where English is not the language of instruction receive between 0.24 and 0.52 points lower median evaluations scores (on a five-point scale) compared to their native-born counterparts, conditional on course type. I also find that being taught by a foreign TA does not have a significant impact on the students' objective course outcomes, such as grades, STEM major declaration, and STEM graduation. These findings suggest that evaluations of teaching for foreign TAs should be used with caution as they might not be a clear reflection of teaching quality.

JEL-Classification: I20, I23, J16, J15
Keywords: Higher education, teaching assistants, STEM persistence

*I would like to thank John Bound, Charlie Brown, Kevin Stange, and Jason Owen-Smith for guidance. I thank Margaret Levenstein, and Jeff Smith, Meera Mahadevan, Justin Wolfers, Mel Stephens, Mike Mueller-Smith, Susan Dynarski, Gaurav Khanna, Sang Teck Oh, Gail Lucasan and Yeliz Kacamak for their feedback. I acknowledge the NSF SMA-1262447 grant for generous research support. This research has been determined exempt from human subjects control under exemption 1 of the 45 CFR 46.101 (b) by the U.M. Institutional Research Board (HUM00085505). All errors are my own.

1 Introduction

Globalization has generated an increase in the number of non US-born graduate students attending American universities [Bound et al., 2009]. Between 1980 and 2015, the number of international graduate students has more than tripled, reaching a record high of 350, 000 students [Zong and Batalova, 2016]. On one side, the demand from abroad for a U.S. graduate degree has grown rapidly due to higher college completion rates in countries like China and India [Gaulé and Piacentini, 2013]. In addition, because of the high transferability of analytical skills, the demand for a U.S. graduate education has been higher for STEM (Science, Technology, Engineering, and Math) degrees [Bound et al., 2009]. On the supply side, large increases in both federal funding for science and public support for graduate education have provided more opportunities to attend graduate school.

These sizable increases in the number of foreign graduate students have, in turn, caused significant increases in the number of foreign teaching assistants (TAs) in American universities. This study analyzes the impact of the increase in the number of foreign TAs on the educational production function at undergraduate level at the large Midwestern university. [1] The TAs are graduate students who hold office hours, teach smaller sections of the course, and grade assignments and exams. While they are less experienced than the senior staff, they may be able to relate better to the undergraduate students since they share more common experiences, being students at the same university. They constitute an important input in university teaching, making up about 15 percent of the post-secondary instructors in the United States [Bureau of Labor Statistics, 2016].[2]

Given the importance of their contribution to the educational production function, several theories have been invoked to justify why TA characteristics matter for the undergraduate students. Among these theories is the shifting standards theory of stereotyping [Biernat et al., 1991] which suggests that peoples' judgments are influenced by relative comparisons among social groups. This theory suggests that lower status groups (e.g. women and minorities) have a harder time demonstrating competence [Foschi, 2000, Basow et al., 2006]. In the context of this paper, I assume that undergraduate students compare foreign TAs with native TAs when making decisions about the effectiveness of teaching. The most common challenges faced by international students are problems with functionality in the English language and problems with adjusting to the American culture

[1] Figure 1 shows the trend for STEM versus no-STEM foreign graduate students at this university over a period of 13 years.

[2] The authors use the low cost of hiring a TA as one of the potential reasons why TAs make up for a relatively large percentage of instructors. According to Bureau of Labor Statistics [2016], the median annual wage for post-secondary teachers in the U.S. was $75, 430, while the mean wage for TAs was $34, 240.

[Andrade, 2006, Trice, 2003] and these two issues appear to be the reasons why undergraduate students might treat foreign TAs differently [Plakans, 1997]. Based on these previous studies, I assume that foreign TAs differ from their native counterparts in two important dimensions: familiarity with the U.S. culture and their level of English proficiency. To disentangle the effects of cultural and linguistic differences, I consider two categories of foreign TAs, based on whether or not English is an official or de-facto language in their country of origin.[3] In the absence of any indicators of the foreign TAs' English proficiency and assimilation in the American culture,[4] this categorization is a good alternative to disentangle the two ways in which foreign TAs are different from their American counterparts.

Following this definition, this study explores the effect of the increase in foreign TAs on the subjective (student evaluations) and also objective (persistence in STEM majors) outcomes of the undergraduate students. I use administrative data from a large public Midwestern institution that contains information on all students (both undergraduate and graduate) and the courses they attended in each semester between Fall 2001 and Winter 2014. This data also contains information on the TAs for each course, which allows me to characterize the TAs based on country of origin, while also controlling for other TA characteristics such as race, gender, and teaching experience.

I only consider STEM courses taught by TAs given both the large increase in STEM foreign graduate students and also the small percentage of U.S. undergraduate students who major in STEM fields [Xie and Killewald, 2012, Xie et al., 2015]. In large introductory STEM courses, TA-led sessions are one of the few opportunities for undergraduate students to receive small group instruction, so it is important to examine the impact of the large increase in foreign TAs on undergraduate student outcomes. In addition, large introductory STEM courses offer the ideal setting of conditional random assignment of TAs. More specifically, TAs are assigned to each section based on scheduling constraints, both personal and departmental. Thus, at the time of making their choices, both the TAs and the undergraduate students only have access to information about the time and the day in the week of the section. This makes it almost impossible for the undergraduate students to select a section based on the TA, since they cannot see the name of the TAs when signing up for courses. In addition to this, I run balancing tests to show that the characteristics of the undergraduate students are independent of the characteristics of the TA teaching the section, which shows that self-selection into sections of the course is not an issue of concern. This conditional random assignment allows me to draw causality conclusions about the foreign TAs.

[3] One caveat to this explanation is the possibility that TAs from English speaking countries might be closer culturally to the native TAs.

[4] Unfortunately, TOEFL (Test of English as a Foreign Language) scores are not available.

I first analyze the student evaluations of teaching (SETs), which are used by most colleges and universities in the U.S. to make decisions about their instructors [Murray, 2005]. These evaluations provide feedback regarding the quality and effectiveness of the instructors [Svinicki and McKeachie, 2010]. In addition to reflecting teaching quality, SETS have also been shown to reflect teaching effectiveness irrelevant factors [Carrell and West, 2010], such as gender, ethnicity and age [Stark and Freishtat, 2014, Andersen and Miller, 1997, Basow, 1995, Cramer and Alexitch, 2000, Worthington, 2002]. In this paper, I investigate whether the SETs are related to the country of origin of the TAs. I find that a foreign TA from a non-English speaking country has a median evaluation score of overall quality of teaching between 0.24 and 0.52 points lower than an American TA. Even though foreign TAs from English speaking countries get lower evaluation scores, these results are not statistically distinguishable from both the evaluations of native TAs, as well as the ones for TAs from non-English speaking countries.[5]

The evaluation of foreign born TAs is likely to be dependent on both their teaching performance, as well as on other factors such as cultural differences, social skills, and discipline. To test for this, I examine additional evaluation questions regarding the TA effort exerted, course environment and undergraduate student's self-reported learning from the course. Again, I find that TAs from countries where English is not the official or de-facto language are penalized on criteria regarding effort exerted and learning-inducing class environment. However, undergraduate student self-reported learning is not significantly different in sections led by native TAs than in sections led by non-native TAs. These results are consistent with Watts and Lynch [1989] who suggested that that undergraduate students might blame foreign TAs for their poor course performance.

To assess whether evaluations reflect cultural discontent rather than poor teaching skills, I investigate the effect of non-U.S. born TAs on more objective student outcomes, such as grades, declaring a STEM major and graduating in STEM. The results indicate that foreign TAs have no effect on the grade the undergraduate students get in the course. In addition, I do not find any detectable impact of being assigned to a foreign TA in an introductory STEM course on either the probability of declaring a STEM major or the probability of graduating in STEM.

I also show that the lack of impact of foreign TAs on objective outcomes is not driven by the lack of impact of TAs on the undergraduate student outcomes. To establish this, I employ a value-added model framework to test the importance of teaching assistants as an input for students' academic outcomes. Using a random effects model akin to Carrell and West [2010] and De Vlieger et al. [2017], I find substantial variation in student performance across TAs, both in the

[5]Because of the small sample size, the estimates have a low precision.

contemporary class and also in a subsequent class. These results suggest that, while TAs have substantial impacts on undergraduate student outcomes, foreign TAs are not systematically different from native TAs regarding teaching effectiveness.

My findings have broad policy implications and inform us on how having a foreign TA impacts the outcomes of undergraduate students. This study suggests using precaution when taking teaching evaluations as an indicator of teacher quality. My results are consistent with the shifting standards model that implies that undergraduate students evaluate foreign TAs based on preexisting negative stereotypes about their competence as teachers.

The remainder of the paper proceeds as follows: Section 2 reviews the previous literature on TA performance. Section 3 reviews the data and presents information about the institutional background of the data. Section 4 reviews the empirical setting. Section 5 presents the main results of the estimation, Section 6 presents extensions of the analysis and Section 7 presents concluding remarks.

2 Existing literature

Most of the previous papers examining the student evaluations of teaching (SETs) study the connection between the gender of the instructor and their rating of teaching effectiveness. Early findings in this literature show mixed results of instructor gender on SETs [Sidanius and Crane, 1989, Basow and Silberg, 1987, Centra and Gaubatz, 2000, Feldman, 1993]. However, the more recent and also more rigorous studies provide consistent evidence of female instructors receiving lower evaluation scores than their male counterparts [Miller and Chamberlin, 2000, Bianchini et al., 2013, Boring, 2017, Boring et al., 2016].[6] Rosen [2017] examines `RateMyProfessors.com` data and finds that female professors receive significantly lower ratings than male professors. In addition, undergraduate students reward the instructors who follow these gender norms [Sprague and Massoni, 2005, Dalmia et al., 2005], and penalize the ones who don't [Andersen and Miller, 1997]. While the emphasis of previous literature on student evaluations is on gender, little is known on how country of origin impacts the evaluation scores.

Very few previous studies have addressed the efficacy of teaching assistants (TAs), and even fewer have examined the impact of foreign TAs on student performance. The earlier papers on this topic find mixed results of the effect of foreign TAs on undergraduate students' outcomes. Jacobs

[6] According to MacNell et al. [2015], undergraduate students often have different expectations of their instructors, based on their gender. Thus, they expect male instructors to have more "masculine" attributes, such as professionalism and objectivity, while they expect the female ones to be more "feminine"as in having warmth and accessibility.

and Friedman [1988] examine data from three mathematics courses and one business course at a major Midwestern university and find that foreign TAs are just as effective as native TAs when assessing the final examination scores. They also find no significant differences in the ratings of the foreign TAs compared to native TAs and attribute this finding to the extensive TA screening that foreign TAs are required to undertake at the university.

In another earlier study, Norris [1991], analyzes data from three University of Wisconsin-Madison courses (one survey course and two Economics courses) and finds that sections led by non-native English speakers received higher grades. Contrary to this finding, Watts and Lynch [1989] examine data from Purdue University and conclude that international TAs have a negative impact on post-course standardized test scores.[7] Furthermore, they find no statistically significant relationship between foreign TAs and undergraduate student grades, which could indicate that the native TAs were teaching more to the test than the international TAs. None of these early studies, however, present a setting of random assignment of TAs and they control for very few student and TA characteristics.[8]

The more recent papers in this area have examined only economics courses, with the most prominent being Borjas [2000]. In this study, 309 undergraduate students in an intermediate microeconomics course at a large public university are surveyed about their introductory economics courses taken and their experiences with the TAs. The questions from the survey were designed to assess English ability and preparation of foreign born TAs for teaching. The findings show that foreign-born TAs have a negative impact on the undergraduate students' grade. However, foreign born TAs that are better prepared than native TAs do not worsen the achievement of the undergraduate students. Given that the surveys were administrated after the undergraduate students received their grades, these results might be driven by the subjectivity of the answers. For example, as Watts and Lynch [1989] suggest that undergraduate students might blame their bad grades on foreign TAs, and thus modify their answers to the survey accordingly. Furthermore, the empirical strategy presented in the research does not take account of any additional undergraduate student or TA characteristics.

Following up on this work, Fleisher et al. [2002] also investigated the influence of foreign-born TAs on undergraduate students and found little adverse effect on the grades in the courses, which the authors argue is a result of the full year of training that the TAs at the university had to undergo.

[7] The test considered was the revised Test of Understanding College Economics which was designed by the American Economic Association to measure the performance of students in introductory economics courses.

[8] Watts and Lynch [1989] only control for student SAT scores and no additional TA characteristics besides being foreign. Norris [1991] controls for TA experience and high course load, but not any undergraduate student characteristics. Jacobs and Friedman [1988] controls for undergraduate students' SAT scores and the TAs' teaching experience.

This explanation is consistent with previous studies that show that training leads to both higher ratings from the undergraduate students [Shannon et al., 1998] and a higher sense of self-efficacy[9] towards teaching [Prieto and Altmaier, 1994]. Furthermore, Fleisher et al. [2002] also found that foreign-born TAs got lower ratings in students' evaluations of teaching. One explanation brought forth by the authors is that the international TAs might provide a less desirable class environment due to the cultural gap between themselves and the American-born undergraduates or differences in teaching style.

Additional TA characteristics, besides country of origin, were also found to be relevant for undergraduate student performance measures. Among these characteristics, the most researched one is gender. The studies that analyzed the impact of gender on the instructor on undergraduate student outcomes found mixed results when examining a variety of outcomes, among which grades, persistence outcomes (i.e., dropping the course, taking additional courses in the same field, majoring in that field), and attaining an advanced degree [Robst et al., 1998, Canes and Rosen, 1995, Rask and Bailey, 2002, Bettinger and Long, 2005, Rothstein, 1995, Price, 2010]. However, the results on gender matching between TAs and undergraduate students were more indicative of role-model effects: female undergraduate students who have a female TA are less likely to drop out of the course, with no overall effect on performance in the class [Butler and Christensen, 2003]. Another strand of the literature found positive impacts of racial/ethnic matching between undergraduate students and instructors [Price, 2010, Lusher et al., 2015, Fairlie et al., 2014].

This study contributes to the literature on student evaluations and foreign TAs by using rich administrative student data from a public Midwestern institution. In comparison with previous studies, I examine a multitude of STEM courses, using a larger sample of undergraduate students. In addition to this, the institutional setting offers a close to random assignment of TAs to course that allows me to draw causal inferences about the impact of these TAs and undergraduate students' outcomes. I also examine a broad range of outcomes of the undergraduate students, which include shorter term ones such as the grades in courses and declaring a major and also longer term ones, such as graduation rates. In addition to the course outcomes of the undergraduate students, this study analyzes outcomes relating to the evaluations of teaching for the TAs, thus bridging the gap between the two existent study areas. Given the large increase in foreign TAs over the past decade and given that this increase is significantly larger in STEM, it is important to analyze the impact of foreign TAs in the context of these large STEM courses that the undergraduate students take.

[9]The term self-efficacy refers to a person's belief in their ability to accomplish a task [Bandura, 1982].

3 Institutional background and data

This section describes the institutional background and data used for my analysis.

3.1 Institutional background

I use administrative student data from a public Midwestern institution, where the main colleges are the College of Arts and Sciences (which has approximately 60% enrollment) and the College of Engineering.[10] Teaching at the university is done on a semester calendar system, with Fall and Winter semesters, followed by two shorter Spring/Summer semesters.

In addition to the primary faculty member in charge of leading the main lectures, most large introductory courses also have a TA involved in the instruction of the course. The majority of the TAs are current graduate students enrolled at the university. There are some rare instances where undergraduate students are also allowed to teach, but I only consider graduate students in my analysis. The TAs responsibilities vary based on the course and the department and they involve a combination of grading assignments, guiding discussion or laboratory sections, assisting with the preparation of course materials or leading study sessions. This study only considers introductory STEM courses that the undergraduate students take in their first two terms of classes. I denote as STEM all the fields thought to contribute to technological innovation [Xie et al., 2015]. Although there are various STEM definitions, I employ the one used by U.S. Immigration and Customs Enforcement (ICE) for allowing special work visas for foreign nationals in STEM fields [Gonzalez and Kuenzi, 2012]. Unlike the STEM definition used by the National Science Foundation (NSF), the ICE definition[11] doesn't include the social sciences. Given that most social sciences recruit graduate students based on very different criteria than the sciences, I believe that using the ICE definition is the better approach. Furthermore, I use the original ICE definition of STEM and disregard additions to the list of STEM degrees in 2011 and 2012 (when fields like psychology, agriculture, etc. were added to the STEM list).[12]

When applying for a TA position in the STEM courses considered, each graduate student specifies their top preferences regarding which courses they would like to teach. These preferences together with the preferences of the faculty of the course are passed on to the person in the department in charge with TA allocations and assignments, which makes the final decision. No screening

[10] The College of Engineering has a separate admission process, but the students in this college can take courses from all the other colleges of the university.

[11] `https://www.ice.gov/doclib/sevis/pdf/nces_cip_codes_rule_09252008.pdf`

[12] Section B offers a list of the courses considered, which are introductory courses in: biology, chemistry, physics, mathematics and engineering.

is involved, but most of the faculty members already know the graduate students (or can ask their adviser about their background). Thus, the faculty members make informed decisions about which TAs would be best suited for their course. The TAs also undergo training prior to their first semester teaching the course or during the first term teaching, depending on the course considered. In addition to this, TAs from undergraduate universities where courses are taught in languages other than English[13] from the College of Arts and Sciences are required to take a college teaching course from the English Language Institute.[14] The departments also provide access to a graduate student mentor, responsible for giving teaching advice and making observations about teaching. TAs are evaluated based on the median score on the teaching evaluations. If a TAs receives a median evaluation score below 3 (on a Likert scale of 1-5) on the question regarding their overall performance (i.e."Overall, the instructor was an excellent teacher."), they receive a warning from the department. If the poor performance is repeated in a subsequent semester, they will no longer be considered for a TA assignment.

Given the various roles TAs can have in teaching, this study considers three possible types of classes: laboratories, discussion sessions and courses which are taught entirely by TAs. Each individual TA has little input in deciding the undergraduate students' grades, and the degree of input the TA has varies slightly by the type of class considered. The laboratories and discussion sessions do not have separate exams, they only have quizzes and laboratory reports, graded solely by the TAs. Since the grade for the course is determined by exams taken in lecture, I match these sections with the grade in the course. Given the large size of these introductory courses, most of the exams are scantron-graded, multiple choice (additional information on the exams is provided in Appendix B). In the rare cases of non-multiple choice exams, the TAs get together after the exams and grade together using an answer key provided by the faculty teaching the lecture. Given these procedures, it is very unlikely that the difference of grades in the sections to be a result of different grading scales. Therefore, using the grade in the course as an outcome should not be viewed as problematic. In addition to this, I consider additional outcomes that would not be influenced by the TAs' ability to influence grades, such as the probability of declaring a STEM major and the likelihood of graduating in STEM.

[13] This requirement is waved for students who have received their undergraduate degree from a U.S. based institution or from an institution outside of the U.S. with curriculum in English.

[14] All TAs from non-English-medium undergraduate universities are also required to submit their TOEFL exam scores prior to applying to the respective graduate program.

3.2 Data

The data contains all undergraduate students taking classes between Fall 2001 and Winter 2014. The administrative data offers detailed information about the students who are attending this public institution, both undergraduate and graduate students. The data cover the basic demographic information and the entire course taking history of each student. The demographic information includes each student's race (i.e. white, black, Hispanic, Asian, and other (Native American, not indicated, Hawaiian and two or more)), gender (binary male/female), state and country of residency.

For undergraduate students, I use financial aid status in the form of need-based grant eligibility as a proxy for parental income.[15] The largest of the need-based financial grants is the federal Pell Grant, a need-based grant that assists low-income students who are attending universities and other accredited secondary institutions. I create a binary Pell grant variable that identifies students who have received one (or more) of the following grants: Pell Grant, Academic Competitiveness Grant (ACG), Supplemental Educational Opportunity Grant (SEOG) or SMART grant.

I have additional data on Advanced Placement (AP) exams and information about the last high school attended by the student. Since the analysis in this paper focuses on STEM outcomes, I only consider science and math AP tests.[16] In addition to AP test scores, I also control for high school grade point average (GPA), recalculated by the university on a 4.0 scale.[17] SAT and ACT test scores are also included, where SAT scores were standardized into ACT scores using the official ACT conversion table[18].

The data also provide information about the courses taken by the undergraduate students each semester. This information contains the course subject and number, the type of course (lecture, discussion session, laboratory, etc.), the number of credits awarded, and the grade obtained in the course. I define a class as a combination of a term (e.g. Fall 2007), course (e.g. Chemistry 101) and lecture.[19] Three dependent variables are used as a measure of students' achievement in a course: the grade in the course, the probability of declaring a STEM major and the probability of graduating

[15]Data on parental education and income acquired from the admission office has too many missing observations (over 40 percent missing for parental income and over 20 percent for parental education) and multiple imputation methods cannot be used due to the non-randomness of the missing data.

[16]The AP tests considered are: Biology, Chemistry, Physics (Physics B, Physics C: Electricity and Magnetism, and Physics C: Mechanics), Computer Science (Computer Science A and Computer Science AB), Statistics, and Calculus (Calculus AB and Calculus BC).

[17]One caveat is that before 2009, the university included only the courses taken in grades 9-11 for calculating the GPA. After 2009, the university considered all high school courses taken for all grades. However, do not believe that this would be a major issue for my analysis given the richness of my data.

[18]The conversion table can be found at `http://www.act.org/aap/concordance/pdf/reference.pdf`.

[19]For large courses, several lectures might be taught in the same term by different professors. However, TAs are only assigned to one course per term.

in STEM. As explained in Section 3.1, the grades considered are the grades in the course taken by the undergraduate student. In the case where the actual section taught by the TA does not have a separate grade, I consider the grade for the course belongs to. I create a binary variable for majoring in a STEM field by using the CIP (Classification of Instructional Programs) codes that identify each major in combination with the STEM definition from the previous section. I use the same method to create a binary variable for graduation with a STEM degree in five years.[20]

I also control for the race and gender of the TA and use the information about each TA's country of permanent residence at the time of submitting their graduate studies application to create a binary foreign TA dummy.[21] I further divide this foreign TA dummy into two categories, based on whether or not they come from a country where English is an official or de-facto language.

In addition to demographic information on TAs, I also have access to data on the student evaluations of teaching (SETs) from Fall 2008 (when online evaluations were introduced) to Winter 2015. For every course the undergraduate students take each semester, they receive an email in the last week of classes with a link to fill out the teaching questionnaires, followed by three reminders. The timing of filling out the evaluations is such that the students evaluate each course before taking the final exam in that course and learning about their grade. Similarly, the TAs do not have access to the teaching questionnaires filled out by the students until the final grades have been released. This "double-blind" procedure insures that TAs do not award grades based on negative evaluations and that the undergraduate students do not rate TAs based on the final exam or their course grade. Furthermore, the evaluations are anonymous and the TAs receive information about their evaluation scores aggregated at section level.[22] Because of the anonymity of the evaluations, I cannot identify the individual characteristics of each student submitting the evaluation, but I can identify average demographic information about the students at section level (from the data on the courses the students take).

The teaching evaluation form contains questions regarding the course and all the instructors that taught the course, as shown in Figure 4. The questions are designated by department (with some being university wide) and type of instructor (primary faculty or TA). Submitting the evaluations is not mandatory and neither is answering every single question on the evaluations.[23] For each question, the student has a choice of five different answers, which the registrar encodes on a Likert

[20] Similar results are obtained when considering a six year graduation rate.

[21] In contrast with my study, the U.S. Census Bureau defines a foreign-born person as a person who is not a citizen of the U.S. but resides in the country, or a naturalized U.S. citizen.

[22] The only exception to this are the student comments, which are not aggregated. Unfortunately, I do not have access to these comments.

[23] Even though this practice might introduce selection issues, it is still an important issue to examine.

scale: Strongly Disagree=1, Disagree=2, Neutral=3, Agree=4, Strongly Agree=5. Given previous research showing that student answers are likely skewed towards either the lower or the higher end, the registrar calculates the median score rather than the mean for each evaluation question and reports it back to the instructors. Section E explains how to calculate the median score for each evaluation question and provides a computational example.

3.3 Summary statistics

To estimate the effect of foreign TAs, I consider introductory STEM courses[24] that undergraduate students take in their first two semesters of college. This assures that the undergraduate students have minimal prior knowledge about the TAs and that this is their first exposure to college courses. I restrict the sample to undergraduates who entered as Freshmen and were registered for classes between Fall 2001 and Winter 2014. This is important because I do not include transfer students, whose course taking behavior might be different due to past experience. To study graduation rates, I further restrict the sample to undergraduate students taking classes before Winter 2010 to allow a 5 year graduation rate for the last cohort of undergraduate students that I observe. The sample considered is restricted to American undergraduate students in order to eliminate role-model type of behavior. Furthermore, I restrict the sample to only introductory STEM courses that are necessary to take to declare a STEM major.

The courses are also divided based on the component of the course taught by the TA: discussion session, laboratory or full course.[25] The descriptive statistics are presented in Table 1 and they show that laboratories have significantly fewer women than discussion sessions. In general, the sample consists of between 39-48 percent female students, depending on the type of section considered. The sample also consists of almost 70 percent white students, about 5 percent black students, 5 percent Hispanics and 15 percent Asian students. The three types of sections that are led by TAs seem to be balanced in terms of race of the undergraduate students and financial aid status. The courses with laboratories have higher average grades and ACT composite scores than the two other type of courses selected. Furthermore, undergraduate students who take courses with labs are more likely to major in STEM and graduate with a STEM major.

Summary statistics for the teaching evaluations sample are presented in Table 2. When examining the summary statistics divided by section type, Table 2 illustrates that female TAs are less likely to teach a full course than a discussion or lab. The mean age of the TA is approximately 25,

[24] A complete list of the courses that I select in my analysis is presented in Appendix C.

[25] At the university considered, Calculus I and Calculus II are courses taught entirely by the TAs.

the international TAs from English speaking countries make up 5 percent of total TAs in discussion sessions, 8 percent in labs and 16 percent of TAs in full courses. This large variability can be explained by the fact that different departments at the university attract graduate students from various parts of the world (for example, the mathematics department has more students from European countries than the engineering department). About 20 percent of TAs are from non-English speaking countries. The TAs teaching a full course are slightly more likely to have taught more courses before than the other TAs. The median evaluation score for the TA being an excellent instructor is about 4 on a 1-5 scale.

The summary statistics for all TAs (both foreign and native) divided by the country of origin and the type of section is shown in Table 3. The first column of Table 3 shows that the India is the country with the largest number of foreign TAs from English-speaking countries, while the majority of TAs from non-English speaking countries come from China. A similar pattern is true for laboratories, as shown in the second column of Table 3. The analysis for full courses from the third column of Table 3 shows that the majority of the international TAs from non-English speaking countries come from China, followed by Japan and South Korea. The majority of foreign TAs from English-speaking countries come from Canada and India. This analysis also shows that the results for TAs from non-English speaking countries might be driven solely by East Asians.

3.4 Allocation of TAs into classes

One of the main issues raised when estimating teacher quality is the potential non-random assignment of undergraduate students to courses which would bias the estimates. However, this issue is not relevant to this study. First, there is a conditionally-random assignment of TAs: the undergraduate students choose which section to enroll in, but they only see the name of the TA after courses start. Thus, the undergraduate students only see the time of the day and the day of the week of the section. In addition, the TAs had no information about the composition of each section before choosing which one to teach. This reduces the potential self-selection of undergraduate students into a section led by a certain TA. Second, my analysis considers only large introductory STEM courses with capped sections. Thus, there is very little room for the undergraduate students to switch among sections or lectures after they learn who their TA will be.

I also use formal tests to analyze the sorting of undergraduate students into classes. A truly random assignment of undergraduate students would imply that all TA characteristics are unrelated to undergraduate student observable and unobservable characteristics. While I cannot directly test for the correlation of TA characteristics with unobservable undergraduate student characteristics, I

can explore the sorting of undergraduate students into classes based on observable characteristics. More specifically, I regress the average undergraduate student pre-assignment characteristics on TA characteristics in each section of each course and jointly testing the equality of means [De Vlieger et al., 2017].[26] I also include term-course-lecture fixed effects (e.g. Fall 2008, Biology 101, Lecture 100) and add time of the class and day of the week of class as controls.

Since the likelihood of having a foreign TA is highly dependent on the STEM field, it is necessary to add course fixed effects in my analysis. One reason for this is that the undergraduate students taking an introductory STEM class in the fall semester might be different than an undergraduate student taking the same class in the winter (or spring) semester, so I also need to account for the semester the course is taken in. In addition to this, I also need to control for undergraduate students taking the same large lecture to make sure that the undergraduate students in the different sections take the same exams and are exposed to the same professor.[27] Furthermore, controlling for the time of the day and day of the week helps remove the possible selection of undergraduate students or TAs who prefer to attend or teach courses early or late during the day or earlier versus later during the week. I cluster the standard errors at the TA level to account for sections being taught by the same instructors over the course of multiple semesters.

Table 4 shows the results of these balancing tests. The first panel of the table contains randomness checks for discussion sessions. Columns (1), (2), (4), (5), (6) and (7) show that the being from both an English and a non-English speaking country are not significantly related to the undergraduate students' pre-assignment characteristics, such as gender, race (except for black students), financial status, ACT composite scores. Columns (3) and (8) show that TAs from non English speaking countries are marginally less likely to teach black students and students from the state where the university is located. This, however, does not represent a big concern since I control for the undergraduate students' race in all the regressions presented in this study. I also test for differences in assignments of TAs from English speaking countries and TAs from non-English speaking countries and cannot reject the null of no difference (p-values of 0.434 and 0.167, respectively).

I perform similar balance tests for laboratories and full courses, shown in the second and third panels of Table 4. For laboratories, black students are marginally less likely to be in discussions lead by non English speaking TAs. When testing for overall differences in assignment to TAs from different countries, I fail to reject the null of no difference (p-value is 0.208). For randomness

[26] An equivalent method is performed by regressing each undergraduate student's characteristics on course-section indicators and testing the null hypothesis that the coefficients on the course-section indicators are equal to zero [Braga et al., 2016].

[27] Since there are no large lectures for the courses where the TAs teach the entire course (Calculus I and II), I only control for the course and the term.

checks for full courses, English speaking non-American TAs are marginally less likely to teach white students and students with higher ACT composite scores, while non English speaking TAs are marginally less likely to teach Pell grant recipients. When testing for differences in assignments of TAs from English speaking countries and TAs from non-English speaking countries, the only case I fail to reject the null of no differences is for Pell grant recipients. For the remainder of this paper, I control for whether the undergraduate students received a Pell grant in all the regressions presented.

All in all, the balance tables confirm that assignment of TAs into sections is not correlated with observable undergraduate student characteristics, which further informs me that I can credibly estimate the causal effect of the characteristics of the TA on undergraduate student outcomes using least squares regressions.[28]

4 Empirical strategy

4.1 Course evaluations

In this section, I study the impact of the country of origin of the TA on student teaching evaluations. I estimate the impact of foreign TAs on four important outcomes: the overall TA rating, the degree of effort the undergraduate students believe the TA exerted, the course environment and the self-reported student learning in the course. As explained in the previous section, the question about the overall quality of the TA[29] is the most important question on the TA evaluation questionnaire and it determines the likelihood of the graduate student receiving a teaching assignment in the future. The distribution of the answers for this question is presented in Figure 2 and shows that most of the evaluation scores are between 4 and 5.

I also consider other evaluation categories in my analysis. One important evaluation category is the degree of effort the undergraduate students believe the TA exerted. As seen from Table 9, these questions relate to how promptly the TA graded assignments, how well they handled questions in the class, how prepared they were for the class, and how knowledgeable they were about the subject taught. In the case that a section contains multiple of these evaluation questions, I take the average of the median answers. Another group of evaluation questions that I consider relates to the course environment (as shown in Table 10). These questions depend greatly on the course considered, and they relate to how fair the TA was, how willing the TA was to help the undergraduate

[28] I can make this claim by assuming that the student characteristics that are not correlated with observable undergraduate student characteristics are also not correlated with observable TA characteristics.

[29] The question varies slightly across the courses considered, as shown in Table 8.

students outside the class, how enthusiastic the TA was, and whether the TA enjoyed teaching the class. Even though these questions do not relate directly to undergraduate student learning or TA preparedness, I believe they are an important factor in determining the perceptions of undergraduate students regarding the TAs and the country of origin of the TAs. One last category I consider is the self-reported undergraduate student learning in the course. Table 11 shows the questions from the evaluation form that were selected to indicate how much the students think they learned from the specific course. All these evaluations questions refer to only the section taught by the TA, and not the course as a whole. I use the following regression to analyze the impact of foreign TAs on median student evaluation scores:

$$\begin{aligned} y_{cst} = \alpha_0 + \alpha_1 X_{cst} + \alpha_2 Z_{cst} + \gamma_1 \text{Engl speaking foreign TA}_{cst} \\ + \gamma_2 \text{Non-Engl speaking foreign TA}_{cst} \\ + \rho_{ct} + \epsilon_{cst} \end{aligned} \tag{1}$$

I define the outcome y_{cst} to be the outcome for section s, in term t, for course c, which is the median score of teaching evaluation for the four categories considered: the overall quality, the degree of effort the undergraduate students believe the TA exerted, the course environment and the self-reported undergraduate student learning in the course. This score is a section level aggregate score calculated by the institution using the formula for finding the median of a grouped frequency distribution (found in Appendix E). The variables of interest are the binary variables indicating a foreign TA from an English speaking county and a foreign TA from a non-English speaking country. The vector X_i contains controls for TA characteristics such as gender, race, age, and Z_{cst} is the vector of controls for the average undergraduate student characteristics in each course c, in section s, in term t. Since evaluations are anonymous, I can only control for average undergraduate student characteristics in the respective sections of the course. I also include term-course-lecture fixed effects (ρ_{ct}) and add time of the class, and day of the week of class as controls. The standard errors are clustered at TA level.

4.2 Undergraduate student course performance

In this section, I present an analogous ordinary least squares model to the one in the previous section, with the scope of analyzing how the TA's country of origin influences undergraduate student outcomes. I employ the following regression model:

$$
\begin{aligned}
y_{itcs} = \beta_0 + \beta_1 X_i + \beta_2 Z_{tcs} + \gamma_1 &\text{Engl speaking foreign TA}_{tcs} \\
&+ \gamma_2 \text{Non-Engl speaking foreign TA}_{tcs} \\
&+ \rho_{ct} + \epsilon_{itcs},
\end{aligned} \tag{2}
$$

where y_{itcs} is the outcome measure for undergraduate student i in course c and section s, in semester t. It should be noted that this model is very similar to the model presented in the previous section, with the difference being that I control for individual undergraduate student characteristics, and not section averages like in the previous analysis. The outcomes considered are the grade in the class, ever having declared a STEM major, and graduating with a STEM degree in 5 years. Since the majority of undergraduate students graduate in 5 years as compared 4 years, I allow undergraduate students to take 5 years to graduate. The model considered includes controls for international TAs, both from English speaking countries as well as non-English speaking countries. Once again, the coefficients of interest are γ_1 and γ_2. X_i are the controls for undergraduate student demographics and course taking behavior (gender, race, ACT composite score, high school GPA, financial aid) and Z_{ics} are the controls for TA characteristics such as gender, race, and age.

Given that each undergraduate student could take multiple introductory STEM courses in the first year and given that these courses could be taught by the same instructors (even though not in the same semester), it is necessary to cluster the standard errors at both the undergraduate student level, as well as at the TA level. Cameron et al. [2011] propose a new variance estimator for OLS that provides cluster-robust inference when there is a two-way clustering that is non-nested. Correia [2016] improves this two and multi-way clustering of standard errors by also allowing for absorption of multiple fixed effects. Therefore, I use the command developed by Correia [2016] to be able to get the correct standard errors for my estimation. Also included in the regression are term-course-lecture fixed effects (ρ_{ct}).

5 Results

5.1 Evaluations

The first panel of Table 5 provides the estimation results for the overall quality of TAs. The results show that TAs from non-English speaking countries get significantly lower median evaluation scores than native TAs, with a median score between 0.24 and 0.52 points lower. This is a relatively large effect of about half of a standard deviation, with the average across the three samples close to

4.

This effect is only about one third of the effect that Fleisher et al. [2002] get, but in their research they do not control for other TA characteristics besides country of origin. The estimated effects for non-American TAs from English speaking countries are also negative, although not statistically significant. Interestingly, although female TAs do get lower median evaluation scores than the male TAs in the courses selected, the results are not statistically significant once I control for other TA characteristics.[30] Furthermore, non-white, non-Asian TAs (i.e. blacks, Hispanics, and other races) are also penalized for evaluation scores, with very large effects for the discussion sessions.

Table 5 also provides the results of a F-test for the equality of coefficients for the TAs from English-speaking countries and TAs from non-English speaking countries. I fail to reject that the impact of a foreign TA from an English-speaking country on the median evaluation score is the same as the impact of having a foreign TA from a non-English speaking country at a 5 percent significance level.

The rest of the panels in Table 5 show the results for the additional evaluation questions considered. The results suggest that foreign TAs from countries that do not have English as their official/de-facto language are perceived as being worse at exerting effort and promoting a desirable class environment. These results are consistent across the different sections considered and significant, except for TA effort in laboratories. These results also show that being a foreign TA from a non-English speaking country lowers the median evaluation score by about half of a standard deviation of the median evaluation scores, where the mean is around 4. Foreign TAs from English speaking countries also get lower evaluation scores as compared with their native counterparts regarding TA effort, but the results are only significant for the courses where they teach the full course.

When estimating the impact of international TAs on the course environment, foreign TAs get lower median evaluation scores and the results are significant for TAs from non-English speaking countries. One last evaluation question that I consider is the one regarding self-reported undergraduate student learning. Except for full courses, none of the results for foreign TAs is statistically significant. This question is also more connected to the results that I present in the next section that involves the undergraduate students' objective outcomes.

Systematically, this set of results show that TAs from non English speaking countries are getting lower median evaluation scores than native TAs on all questions considered except for the ones about undergraduate student learning. The next step is to examine the impact of TA country of

[30]This result is different from Boring [2017] who finds that students believe that women have a comparative advantage in course preparation and organization of courses, while men have a comparative advantage in class leadership skills.

origin for both short-term and long-term student objective student outcomes.

5.2 Course grade

This section provides the estimation results using the model from the previous subsection. I study the impact of the having a foreign TA on both short-term student outcomes and long-term ones. Table 6 shows the estimation results for the ordinary least squares model that has the grade received in the course as the outcome. Each course has letter grades A-E, which are converted to the standard 0-4 scale.[31] I present the results for the three types of TA-led sections that I consider in my analysis. The estimated effect of TAs from non-English speaking countries from Table 6 is negative, small and insignificant. The point estimate indicates that having a TA from a non-English speaking country reduces the grade by 0.03-0.04 points, which is one tenth of the difference from a grade to the next one (e.g from B to B+), and it's only around 5 percent of a standard deviation of the grade variable, with a mean of about 3. Besides this effect not being significant, it also constitutes only around one sixth of the effect of one point change in the ACT composite score on the grade in the course. The results indicate that having a TA from a non-English speaking country reduces the grade in the course by 3-4 percent of standard deviation. Even though not directly comparable, these results are slightly lower than the previous results found in the literature, where Lusher et al. [2015] find that undergraduate students' grades increase between 2 and 4 percent when exposed to TAs of their own ethnicity.

5.3 Other outcomes

One concern is that contemporary course grades do not fully capture the full TA effectiveness [Jackson, 2013] and they are just a reflection of different grading policies or standards across TAs. I address this issue by investigating whether having a foreign TAs impacts the undergraduate students' ability for deep learning, a concept used by Carrell and West [2010] to refer to persistent effects of undergraduate student learning. I quantify the effects of deep learning by considering the probability that an undergraduate student ever declared a STEM major and the probability that the undergraduate student graduated with a STEM degree in 5 years. Studying these additional outcomes also addresses any concerns of the TAs having any input on the course grades.

The results using STEM declaration as an outcome are shown in the second panel of Table 6. The estimation results show that undergraduate students who have a non English speaking TA in

[31] A+, A =4.0 points, A-=3.7, B+ =3.3, B =3.0, B- =2.7, C+ =2.3, C=2.0, C-=1.7, D+=1.3, D=1.0, D-=0.7 and E=0.0

discussion sessions have a slightly higher probability of declaring a STEM major. More specifically, in discussion sessions, having a foreign TA from non-English speaking country increases the probability of majoring in STEM by about 3 percentage points relative to a mean of 60 percent, which corresponds to about 5 percent difference. I am also interested in longer-term outcomes, such as college STEM graduation. The results for five-year graduation rates are shown in the last panel of Table 6. The point estimates for country of origin of TA are again very tiny and they indicate no effect of foreign TAs on the undergraduate students' deep learning. All in all, these results indicate that there is no clear evidence that foreign TAs are doing any worse than native TAs in terms of teaching effectiveness, as measured by actual undergraduate student outcomes.

Once again, I also perform F-tests to test whether the impact of having a foreign TA from a non-English speaking country on objective student outcomes is the same as the impact of a TA from an English-speaking country on the same outcomes. For all the three different outcomes considered (grades, probability of declaring a STEM major and probability of graduating with a STEM degree), I find that I cannot reject the equality hypothesis at the 5 percent level.

6 Extensions

6.1 Robustness checks

I consider the sensitivity of my results to the inclusion of different controls. Table 18 presents these results. The first column of the table shows the regression results including both the undergraduate student and TA controls, the second column only includes TA controls, and the last column only includes undergraduate student controls. I present robustness checks only for two of the outcomes considered: median evaluation scores for the overall teaching effectiveness and grades in the course. Across the different specifications considered, we can see that the results are robust to the exclusion of different controls, with the coefficient estimates changing the most when not including TA controls.

6.2 Does TA quality matter?

The previous results could be explained by the fact that perhaps TAs don't really affect grades. One method to evaluate the TAs based on their impact on the undergraduate students' grades is the value-added (VA) approach, first implemented by [Hanushek, 1971] and [Murnane, 1975].

The majority of studies relying on the value-added framework have been written in the context of primary and secondary schools [Rockoff, 2004, Rivkin et al., 2005, Chetty et al., 2014a,b,

Rothstein, 2010, Hanushek, 1971, Kane and Staiger, 2008]. A handful of studies have looked at the variation of professor effectiveness at the university level and found that instructor effectiveness explains a significant share of the variation in undergraduate students' grades [De Vlieger et al., 2017, Carrell and West, 2010, Brodaty and Gurgand, 2016], subsequent courses [Bettinger and Long, 2010, Figlio et al., 2015, Carrell and West, 2010] and labor market outcomes [Braga et al., 2016].

In this section, I provide evidence of the existence of variation in TA effectiveness. I consider the same sample of undergraduate students as in my previous analysis taking two two introductory STEM courses: Calculus I and Calculus II. As explained in Appendix C, the exams in Calculus I and II are not multiple choice, but the TAs have very little room for influencing the undergraduate students' grades as the exams are uniform among all sections of the course and the TAs get together to grade (a group of TAs are assigned the same question to grade for all the exams).

Even though value-added modeling (VAM) is an important tool used by researchers, there are conflicting conclusions on the degree of bias and instability of the VAMs [Kane and Staiger, 2008, Rothstein, 2010]. One potential factor that could bias the value-added model is the non-random sorting of undergraduate students [Koedel et al., 2015]. Given this concern, balance test were performed (not shown) to assess students' sorting into sections.

I implement my analysis on TA effectiveness in two steps, by using a random effects model similar to the one used by Carrell and West [2010] and De Vlieger et al. [2017]. The first step involves estimating the following value-added model using ordinary least-squares:

$$Y_{ijkt} = \beta_1 X_i + \beta_2 Z_{jkt} + \gamma_t + \theta_k + \epsilon_{ijkt}, \tag{3}$$

where I define Y_{ijkt} as the outcome of student i in section j taught by TA k during term t. Here, X_i is the vector of undergraduate student characteristics, Z_{jkt} is the vector of section mean peer characteristics. The regression further controls for unobserved differences in academic achievement across time and grade inflation (γ_t). The coefficient of interest is θ_k, which represents the TA value-added or the contribution of TA k to the performance of the undergraduate students. More specifically, I am interested in the variance of θs across TAs, which measures the dispersion of TA quality. The corresponding distribution of TA fixed effects is presented in Figure 3 and it suggests a large variability in TA effectiveness across the different TAs considered.

The second step is to construct average residuals for each section for each outcome:

$$\tilde{Y}_{jkt} = \sum_{i \in j} \left(Y_{jkt} - \hat{\beta}_1 X_i - \hat{\beta}_2 Z_{jkt} - \hat{\gamma}_t - \hat{\epsilon}_{ijkt} \right) \tag{4}$$

The two outcomes I consider are contemporaneous grades (grades in Calculus I) and grades in the follow-up course (the grades in Calculus II of the undergraduate students who took Calculus I). I use the mean residuals to estimate the variance of the TA effects θ_k as random effects with maximum likelihood (using the "mixed" command in STATA with unrestricted covariance matrix).[32]

When modeling the error term in equation 3, I assume it is composed of two additive and independent components: a purely random term and a section specific term: $\epsilon_{jkt} = \mu_{jkt} + e_{jkt}$. The section-specific random effects measures common shocks to all undergraduate students in each section, but not common to all classes taught by the same TA. This term is also reflecting the fact that undergraduate students who receive good grades in Calculus I are more likely to receive good grades in Calculus II.[33] Given the two outcomes considered, the grade in Calculus I and the grade in Calculus II, the error terms can be rewritten as:

$$\begin{bmatrix} \epsilon_{jkt}^{\text{Calc I}} \\ \epsilon_{jkt}^{\text{Calc II}} \end{bmatrix} = \begin{bmatrix} \mu_{jkt}^{\text{Calc I}} + e_{jkt}^{\text{Calc I}} \\ \mu_{jkt}^{\text{Calc I}} + \mu_{jkt}^{\text{Calc II}} + e_{jkt}^{\text{Calc II}} \end{bmatrix} \tag{5}$$

where Calc I and Calc II indicate having taken the respective courses.

Based on this, Equation 4 becomes:

$$\begin{bmatrix} \tilde{Y}_{jkt}^{\text{Calc I}} \\ \tilde{Y}_{jkt}^{\text{Calc II}} \end{bmatrix} = \begin{bmatrix} \theta_k^{\text{Calc I}} + \mu_{jkt}^{\text{Calc I}} + e_{jkt}^{\text{Calc I}} \\ \theta_k^{\text{Calc I}} + \theta_k^{\text{Calc I}} + \mu_{jkt}^{\text{Calc II}} + \mu_{jkt}^{\text{Calc II}} + e_{jkt}^{\text{Calc II}} \end{bmatrix} \tag{6}$$

The key parameters of interest are the estimates of variances and correlations of Calculus I TA effects for the grades in both Calculus I and Calculus II, which are: $\text{SD}(\theta_k^{\text{Calc I}})$, $\text{SD}(\theta_k^{\text{Calc II}})$ and $\text{Corr}(\theta_k^{\text{Calc I}}, \theta_k^{\text{Calc II}})$. Table 7 reports the main estimates of the variances and correlations of Calculus I TA effects for grade outcomes. A one-standard deviation increase in Calculus I TA quality is associated with 0.14 and 0.13 standard deviation increase in undergraduate student course grades in Calculus I and Calculus II, respectively. Converted to course grade points, this is about half of a grade step (going from A- to A). These results are slightly larger than the results of Carrell and West [2010] (who find 0.05 and 0.13 for the variances) and slightly smaller than the results of

[32] Teacher effects are modeled as random effects in Corcoran et al. [2011], Konstantopoulos and Chung [2011], Nye et al. [2004] and Papay [2011]. Random effects models are employed to produce empirical Bayes shrinkage estimators, which are more stable than the unshrunken fixed effects models.

[33] Both De Vlieger et al. [2017] and Carrell and West [2010] assume these common shocks by noting that the estimates of $\text{Corr}(\theta_k^{\text{Calc I}}, \theta_k^{\text{Calc II}})$ would be biased in the absence of this assumption.

De Vlieger et al. [2017] (which are 0.30 and 0.20).

Nonetheless, this substantial variation in TA effectiveness both in the current course and also the subsequent course, suggest that TAs do indeed influence undergraduate students' grades and suggest that prior results in this study cannot be explained by the fact that TAs do not make a difference for undergraduate student outcomes, but by the fact that the country of origin of TAs does not make a difference on the undergraduate students' objective outcomes.

7 Conclusion

The goal of this paper is to shed light on the effectiveness of foreign TAs in the education production function by examining both subjective and objective student outcomes. I examine the impact of international TAs in large introductory STEM courses, where TAs are conditionally-randomly assigned to sections. This study concludes that foreign TAs are different than native TAs on two important aspects: lacking knowledge of U.S. culture and institutions and worse English language skills. To distinguish between these two effects, I divide the foreign TAs based on the official language spoken in their home country. My study finds that foreign TAs from non-English speaking countries receive systematically lower evaluation scores than native TAs. However, I find no evidence that these differences translate into differences in grades. Furthermore, when examining longer term outcomes, such as declaring a STEM major and graduating in STEM, I find no evidence that international TAs are detrimental to undergraduate students' measures of deep learning.

My findings have several implications. First, teaching evaluations should be used with caution as they might not be a clear reflection of teacher quality. These findings support previous findings on student evaluations only being weakly correlated to actual teacher quality [Krautmann and Sander, 1999, Weinberg et al., 2009, Carrell and West, 2010, Braga et al., 2014]. Second, the fact that foreign TAs receive lower evaluation scores is problematic because it might limit their ability to find an academic job in the future. More research needs to be done on quantifying the actual impact of scores of teaching evaluations on job prospects of international graduate students. In addition to this, international students might be forced to allocate more of their resources towards teaching and away from research so as to increase their evaluation scores.

Another concern, brought up by Mengel et al. [2017] in the context of gender biased evaluations, is the impact of teaching evaluations on the students' confidence. This impact could be driven by stereotype threat, a situation in which the performance of individuals who belong to a negatively stereotyped group is inhibited. Previous literature shows that students with certain immigrant background underachieve in school [Weber et al., 2015]. In the setting of higher education,

the low teaching evaluations scores received by foreign TAs might hinder their ability to teach well in the subsequent semesters. Furthermore, this negative feedback received from undergraduate students might not only affect the foreign TAs' ability and teaching opportunities, but also their interest in an academic job.

All in all, results inform university policy on the existent biases in the student community. In the U.S., as Boring [2017] notes, student evaluations have two main goals: provide feedback on instructional input and help make decisions regarding hiring, firing or promoting instructors. While evaluations could provide some feedback regarding the effectiveness of instructors, the possible existent biases make them unsuitable to be used as "objective" measures of evaluation of instructors.

References

K. Andersen and E.D. Miller. Gender and student evaluations of teaching. *Political Science and Politics*, 30:216–218, 1997.

M.S. Andrade. International students in English-speaking universities: Adjustment factors. *Journal of Research in International education*, 5(2):131–154, 2006.

A. Bandura. Self-efficacy mechanism in human agency. *American Psychologist*, 37(2):122–147, 1982.

S.A. Basow. Student evaluations of college professors: When gender matters. *Journal of Educational Psychology*, 87(4):656–665, 1995.

S.A. Basow and N.T. Silberg. Student evaluations of college professors: Are female and male professors rated differently? *Journal of Educational Psychology*, 79:308–314, 1987.

S.A. Basow, J.E. Phelan, and L. Capotosto. Gender patterns in college students' choices of their best and worst professors. *Psychology of Women Quarterly*, 30(1):25–35, 2006.

E.P. Bettinger and B.T. Long. Do faculty serve as role models? The impact of instructor gender on female students. *American Economic Review*, 95(2):152–157, 2005.

E.P. Bettinger and B.T. Long. Does cheaper mean better? The impact of using adjunct instructors on student outcomes. *The Review of Economics and Statistics*, 92(3):598–613, 2010.

S. Bianchini, F. Lissoni, and Pezzoni. M. Instructor characteristics and students' evaluations of teaching effectiveness. *European Journal of Engineering Education*, 38(1):38–57, 2013.

M. Biernat, M. Manis, and T.E. Nelson. Stereotypes and standards of judgment. *Journal of Personality and Social Psychology*, 60(4):485–499, 1991.

A. Boring. Gender biases in student evaluations of teaching. *Journal of Public Economics*, 145: 27–41, 2017.

A. Boring, K. Ottoboni, and P. Stark. Student evaluations of teaching (mostly) do not measure teaching effectiveness, 2016. URL `https://www.scienceopen.com/document?vid=818d8ec0-5908-47d8-86b4-5dc38f04b23e`.

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

J. Bound, S. Turner, and P. Walsh. Internationalization of US doctorate education. In *Science and Engineering Careers in the United States: An Analysis of Markets and Employment*, pages 59–97. University of Chicago Press, 2009.

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

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

T. Brodaty and M. Gurgand. Good peers or good teachers? Evidence from a French university. *Economics of Education Review*, 54:62–78, 2016.

U.S. Department of Labor Bureau of Labor Statistics. Occupational Employment Statistics, 2016. URL `www.bls.gov/oes/`.

D.M. Butler and R. Christensen. Mixing and matching: The effect on student performance of teaching assistants of the same gender. *Political Science and Politics*, 36(4):781–786, 2003.

A.C. Cameron, J.B. Gelbach, and D.L. Miller. Robust inference with multiway clustering. *Journal of Business and Economic Statistics*, 29(2):238–249, 2011.

B.J. Canes and H.S. Rosen. Following in her footsteps? Women's choices of college majors and faculty gender composition. *Industrial and Labor Relations Review*, 48(3):486–504, 1995.

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

J.A. Centra and N.B. Gaubatz. Is there gender bias in student evaluations of teaching? *The Journal of Higher Education*, 70:17–33, 2000.

R. Chetty, J.N. Friedman, and J.E. Rockoff. Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *The American Economic Review*, 104(9):2593–2632, 2014a.

R. Chetty, J.N. Friedman, and J.E. Rockoff. Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *The American Economic Review*, 104(9):2633–2679, 2014b.

S.P. Corcoran, J.L. Jennings, and A.A. Beveridge. Teacher effectiveness on high-and low-stakes tests. *Society for Research on Educational Effectiveness*, 2011.

S. Correia. A feasible estimator for linear models with multi-way fixed effects, 2016. URL `http://scorreia.com/research/hdfe.pdf`.

K.M. Cramer and L.R. Alexitch. Student evaluations of college professors: identifying sources of bias. *Canadian Journal of Higher Education*, 30(2):143–164, 2000.

S. Dalmia, D.C. Giedeman, H. A. Klein, and N.M. Levenburg. Women in academia: An analysis of their expectations, performance and pay. *Forum on Public Policy*, 1:160–177, 2005.

P. De Vlieger, B. Jacob, and K. Stange. Measuring instructor effectiveness in higher education. In C.M. Hoxby and K. Stange, editors, *Productivity in Higher Education*. University of Chicago Press, 2017.

R.W. Fairlie, F. Hoffmann, and P. Oreopoulos. A community college instructor like me: Race and ethnicity interactions in the classroom. *The American Economic Review*, 104(8):2567–2591, 2014.

K.A. Feldman. College students' views of male and female college teachers: Part II. Evidence from students' evaluations of their classroom teachers. *Research in Higher Education*, 34:151–211, 1993.

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

B. Fleisher, M. Hashimoto, and B. Weinberg. Foreign GTAs can be effective teachers of Economics. *The Journal of Economic Education*, 33(4):299–325, 2002.

M. Foschi. Double standards for competence: Theory and research. *Annual Review of Sociology*, 26:21–42, 2000.

P. Gaulé and M. Piacentini. Chinese graduate students and US scientific productivity. *Review of Economics and Statistics*, 95(2):698–701, 2013.

H.B. Gonzalez and J.J. Kuenzi. Science, technology, engineering, and mathematics (STEM) education: a primer. *Congressional Research Service*, 2012. URL `http://www.stemedcoalition.org/wp-content/uploads/2010/05/STEM-Education-Primer.pdf`. Washington, D.C.

E. Hanushek. Teacher characteristics and gains in student achievement: Estimation using micro data. *The American Economic Review*, 61(2):280–288, 1971.

W.L. Hays. *Statistics for the Social Sciences*. New York: Holt, Rinehart and Winston, 2nd edition, 1973.

C.K. Jackson. Non-cognitive ability, test scores, and teacher quality: Evidence from 9th grade teachers in North Carolina. Technical report, NBER Working Paper No. 18624, 2013.

L.C. Jacobs and C.B. Friedman. Student achievement under foreign teaching associates compared with native teaching associates. *The Journal of Higher Education*, 59(5):551–563, 1988.

T.J. Kane and D.O. Staiger. Estimating teacher impacts on student achievement: An experimental evaluation. Technical report, National Bureau of Economic Research, 2008. NBER Working Paper 1460.

C. Koedel, K. Mihaly, and J.E. Rockoff. Value-added modeling: A review. *Economics of Education Review*, 47:180–195, 2015.

S. Konstantopoulos and V. Chung. The persistence of teacher effects in elementary grades. *American Educational Research Journal*, 48(2):361–386, 2011.

A.C. Krautmann and W. Sander. Grades and student evaluations of teachers. *Economics of Education Review*, 18(1):59–63, 1999.

L. Lusher, D. Campbell, and S. Carrell. TAs like me: Racial interactions between graduate teaching assistants and undergraduates. Technical report, National Bureau of Economic Research, 2015.

L. MacNell, A. Driscoll, and A.N. Hunt. What's in a name: exposing gender bias in student ratings of teaching. *Innovative Higher Education*, 40(4):291–303, 2015.

F. Mengel, J. Sauermann, and U. Zölitz. Gender bias in teaching evaluations. IZA Discussion Paper No. 11000, 2017. URL `ftp://repec.iza.org/RePEc/Discussionpaper/dp11000.pdf`.

J. Miller and M. Chamberlin. Women are teachers, men are professors: A study of student perceptions. *Teaching Sociology*, 28(4.):283–298, 2000.

R.J. Murnane. *The impact of school resources on the learning of inner city children*. Cambridge, MA: Ballinger Publishing, 1975.

H.G. Murray. Student evaluation of teaching: Has it made a difference? In *Annual Meeting of the Society for Teaching and Learning in Higher Education. Charlottetown, Prince Edward Island*, 2005.

T. Norris. Nonnative English-speaking teaching assistants and student performance. *Research in Higher Education*, 32(4):433–448, 1991.

B. Nye, S. Konstantopoulos, and L.V. Hedges. How large are teacher effects? *Educational evaluation and policy analysis*, 26(3):237–257, 2004.

J.P. Papay. Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal*, 48(1):163–193, 2011.

B.S. Plakans. Undergraduates' experiences with and attitudes toward international teaching assistants. *TESOL quarterly*, 31(1):95–119, 1997.

J. Price. The effect of instructor race and gender on student persistence in STEM fields. *Economics of Education Review*, 29(6):901–910, 2010.

L. Prieto and E. Altmaier. The relationship of prior training and previous teaching experience to self-efficacy among graduate teaching assistants. *Research in Higher Education*, 35(4):481–497, 1994.

K.N. Rask and E.M. Bailey. Are faculty role models? Evidence from major choice in an undergraduate institution. *The Journal of Economic Education*, 33(2):99–124, 2002.

S.G. Rivkin, E.A. Hanushek, and J.F. Kain. Teachers, schools, and academic achievement. *Econometrica*, 73(2):417–458, 2005.

J. Robst, J. Keil, and D. Russo. The effect of gender composition of faculty on student retention. *Economics of Education Review*, 17(4):429–439, 1998.

J.E. Rockoff. The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review*, 94(2):247–252, 2004.

A.S. Rosen. Correlations, trends and potential biases among publicly accessible web-based student evaluations of teaching: a large-scale study of RateMyProfessors.com data. *Assessment & Evaluation in Higher Education*, pages 1–14, 2017.

D.S. Rothstein. Do female faculty influence female students' educational and labor market attainments? *Industrial and Labor Relations Review*, 48(3):515–530, 1995.

J. Rothstein. Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics*, 125(1):175–214, 2010.

D. Shannon, D. Twale, and M. Moore. TA teaching effectiveness: The impact of training and teaching experience. *The Journal of Higher Education*, 69(4):440–466, 1998.

J. Sidanius and M. Crane. Job evaluation and gender: The case of university faculty. *Journal of Applied Social Psychology*, 19:174–197, 1989.

J. Sprague and K. Massoni. Student evaluations and gendered expectations: What we can't count can hurt us. *Sex Roles*, 53:779–793, 2005.

P.B. Stark and R. Freishtat. An evaluation of course evaluations. Science Direct, 2014. URL https://www.scienceopen.com/document?vid=42e6aae5-246b-4900-8015-dc99b467b6e4.

M. Svinicki and W.J. McKeachie. *McKeachie's teaching tips: Strategies, research, and theory for college and university teachers*. Belmont, CA: Wadsworth., 13 edition, 2010.

A.G. Trice. Faculty perceptions of graduate international students: The benefits and challenges. *Journal of Studies in International Education*, 7(4):379–403, 2003.

M. Watts and G.J. Lynch. The principles courses revisited. *The American Economic Review*, 79 (2):236–241, 1989.

S. Weber, M. Appel, and N. Kronberger. Stereotype threat and the cognitive performance of adolescent immigrants: The role of cultural identity strength. *Contemporary Educational Psychology*, 42:71–81, 2015.

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

A.C. Worthington. The impact of student perceptions and characteristics on teaching evaluations: A case study in finance education. *Assessment and Evaluation in Higher Education*, 27(1): 49–64, 2002.

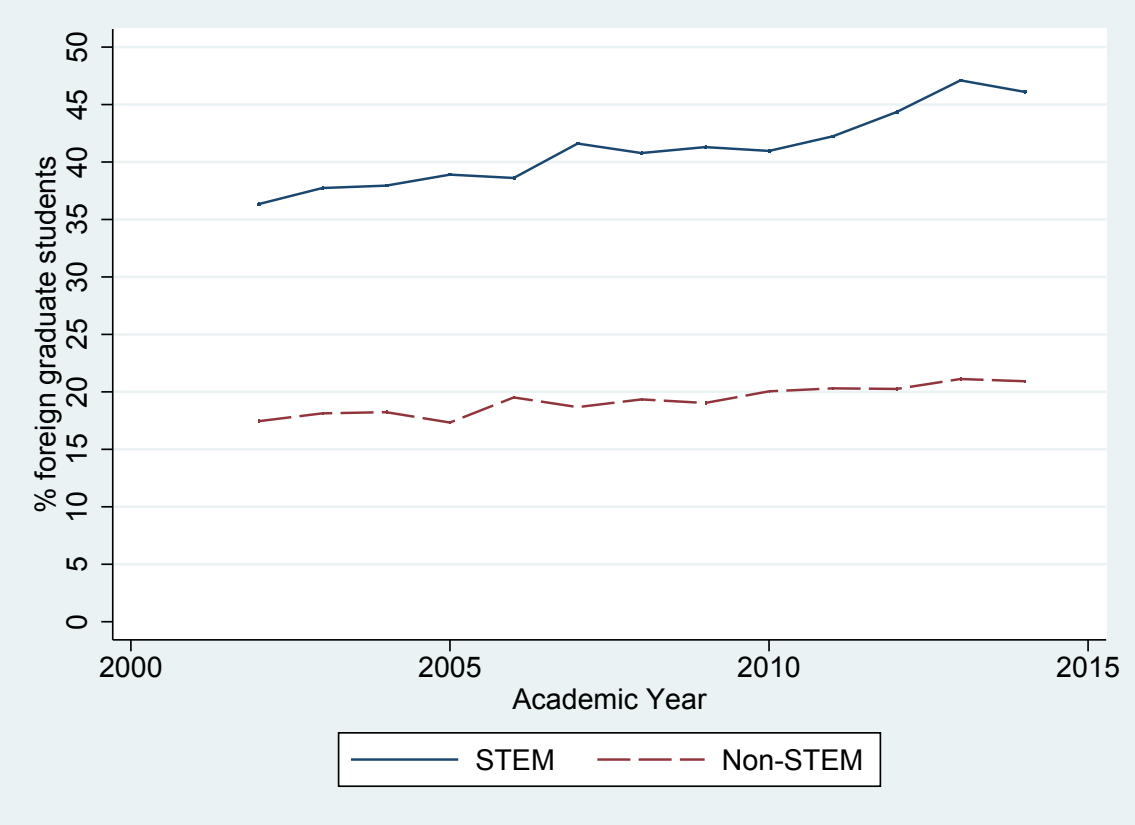
Y. Xie and A.A. Killewald. *Is American science in decline?* Cambridge: Harvard University Press., 2012.

Y. Xie, M. Fang, and K. Shauman. STEM education. *Annual Review of Sociology*, 41:331–357, 2015.

J. Zong and J. Batalova. International students in the United States. Migration Policy Institute, 2016. URL http://www.migrationpolicy.org/article/international-students-united-states.

A Tables and Figures

Figure 1: Share of foreign graduate students in STEM and non-STEM programs at a large public Midwestern university



Notes: The figure shows the share of foreign graduate students in STEM and non-STEM programs at a large public Midwestern university over 2001-2014.

Table 1: Summary statistics for outcomes

	Discussion		Laboratory		Full course	
	Mean	SD	Mean	SD	Mean	SD
Female	0.48	0.50	0.39	0.48	0.40	0.49
White	0.66	0.47	0.68	0.47	0.70	0.46
Black	0.06	0.23	0.04	0.20	0.04	0.20
Hispanic	0.05	0.21	0.04	0.20	0.05	0.22
Asian	0.16	0.36	0.16	0.36	0.13	0.33
Other race	0.04	0.19	0.04	0.19	0.04	0.18
Pell grant	0.20	0.40	0.19	0.39	0.20	0.40
ACT composite score	28.80	3.09	29.22	3.04	28.67	2.74
In state	0.75	0.43	0.73	0.44	0.70	0.46
HS GPA	3.79	0.23	3.82	3.80	3.77	0.24
HS GPA Missing	0.08	0.27	0.07	0.25	0.08	0.28
Grade course	2.80	0.90	3.09	0.82	2.59	0.99
Declared STEM major	0.53	0.50	0.69	0.46	0.51	0.50
Ever graduated with STEM degree	0.42	0.49	0.55	0.50	0.39	0.49
Unique undergraduate students	15256		13957		7729	

Table 2: Summary statistics for evaluations

	Discussion		Laboratory		Full course	
	Mean	SD	Mean	SD	Mean	SD
Female TA	0.50	0.50	0.41	0.49	0.25	0.43
White TA	0.62	0.48	0.51	0.50	0.57	0.50
Black TA	0.014	0.12	0.032	0.18	0.013	0.11
Hispanic TA	0.056	0.23	0.074	0.26	0.063	0.24
Asian TA	0.24	0.43	0.32	0.46	0.27	0.45
Other race TA	0.028	0.16	0.047	0.21	0.030	0.17
Age	25.1	2.54	25.7	3.52	24.7	2.13
Foreign TA from English speaking country)	0.056	0.23	0.083	0.28	0.16	0.36
Foreign TA from non-English speaking country	0.18	0.38	0.25	0.44	0.23	0.42
Times taught	4.19	2.37	5.21	2.97	5.76	2.50
Median evaluation score	3.95	0.72	4.05	0.74	4.08	0.75
Number of sections	761		822		300	
Number of unique TAs	191		303		148	

Table 3: TAs distribution by country of origin

Countries	Number of TAs		
	Discussion sessions	Laboratories	Full courses
English-speaking countries			
Australia	0	2	3
Canada	2	5	6
Ghana	0	1	0
Hong Kong (China)	0	1	1
India	6	6	6
Israel	0	1	0
Jamaica	0	2	0
Malaysia	0	1	1
Singapore	1	0	2
South Africa	0	1	1
Trinidad & Tobago	1	0	0
United States	142	211	83
Non-English-speaking countries			
Argentina	1	1	0
Brazil	0	1	1
Chile	0	1	1
China	28	51	24
Costa Rica	1	0	0
Colombia	0	1	1
Ecuador	1	0	0
Egypt	0	1	0
Greece	0	1	1
Hungary	1	0	0
Iran	0	2	1
Japan	3	0	0
Mexico	1	0	1
Panama	1	1	0
Peru	0	1	1
Romania	0	0	1
Russia	0	0	2
South Korea	3	6	7
Sri Lanka	0	1	0
Sweden	0	0	1
Taiwan	0	1	2
Thailand	0	1	
Vietnam	0	1	1
Total	191	303	148

Table 4: Balancing test of TAs on undergraduate student characteristics for discussion sessions

	Outcomes							
VARIABLES	Avg. female	Avg. white	Avg. black	Avg. Hisp.	Avg. Asian	Avg. Pell	Avg. ACT comp.	Avg. in-state
Discussion sessions								
Foreign TA from non-English speaking country	0.012	0.007	-0.014*	0.000	0.011	-0.017	0.144	-0.029**
	(0.019)	(0.016)	(0.008)	(0.005)	(0.012)	(0.015)	(0.128)	(0.012)
Foreign TA from English speaking country	0.019	-0.033	0.025	-0.000	0.017	0.042	-0.541	-0.023
	(0.022)	(0.039)	(0.027)	(0.008)	(0.019)	(0.046)	(0.412)	(0.025)
F-Test p-value	[0.6786]	[0.8190]	[0.4344]	[0.5736]	[0.9540]	[0.1367]	[0.1397]	[0.1679]
Laboratories								
Foreign TA from non-English speaking country	-0.007	0.001	-0.013*	0.001	0.011	0.016	0.080	-0.019
	(0.012)	(0.018)	(0.007)	(0.006)	(0.012)	(0.014)	(0.098)	(0.013)
Foreign TA from English speaking country	-0.008	-0.008	-0.007	-0.005	0.010	0.013	-0.036	0.002
	(0.017)	(0.026)	(0.006)	(0.008)	(0.016)	(0.016)	(0.120)	(0.021)
F-Test p-value	[0.1706]	[0.9954]	[0.7832]	[0.2064]	[0.5259]	[0.6098]	[0.2447]	[0.2897]
Full courses								
Foreign TA from non-English speaking country	0.001	-0.010	-0.001	-0.008	-0.002	-0.031**	0.182**	-0.013
	(0.019)	(0.022)	(0.006)	(0.006)	(0.014)	(0.016)	(0.083)	(0.016)
Foreign TA from English speaking country	0.018	-0.026*	0.013*	0.005	0.015	-0.002	-0.199**	0.024
	(0.014)	(0.013)	(0.007)	(0.008)	(0.010)	(0.011)	(0.080)	(0.018)
F-Test p-value	[0.4482]	[0.1118]	[0.2081]	[0.5819]	[0.0016]	[0.0589]	[0.3081]	[0.1438]

Notes: Each column is a regression of section-level undergraduate student average characteristics on TA characteristics. We control for foreign TA from a non-English speaking country, as well as foreign TA from English speaking country. Additional controls include the race of the TA, gender, race, teaching experience, and age. All specifications include course-term fixed effects. The robust standard errors are clustered by TA.
The discussion sample size is 761, the laboratories sample size is 822, and the full courses sample size is 300.

Figure 2: Distribution of median evaluation scores

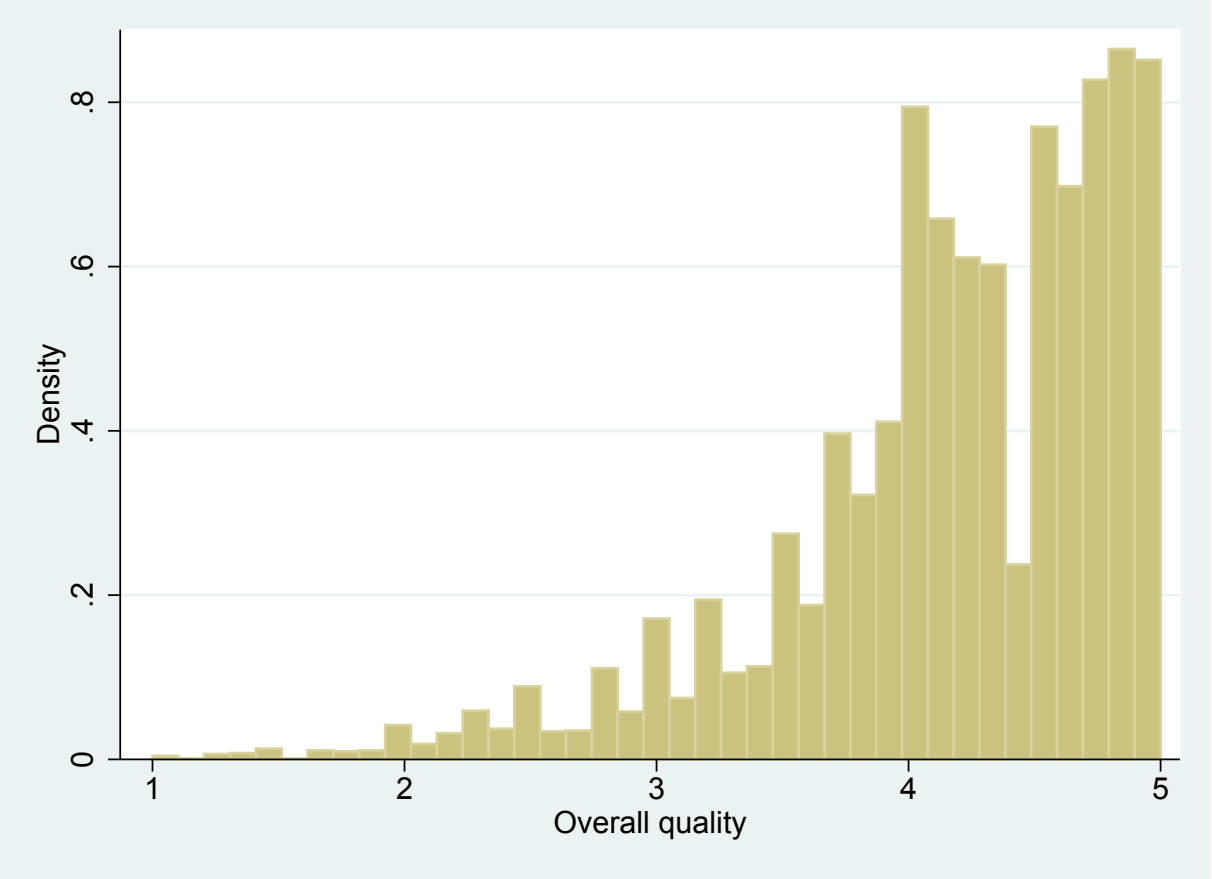


Table 5: Results for median evaluation scores (OLS regression models)

VARIABLES	(1) Discussion	(2) Laboratory	(3) Full course
Overall quality of TAs			
Foreign TA from non-English speaking country	-0.36**	-0.24*	-0.52***
	(0.13)	(0.10)	(0.16)
Foreign TA from English speaking country	-0.35	-0.05	-0.27
	(0.19)	(0.21)	(0.15)
F-test for equality of coefficients	0.92	0.34	0.14
Mean dep. var.	3.95	4.05	4.08
SD dep. var.	0.72	0.74	0.76
Observations	763	822	300
TA effort			
Foreign TA from non-English speaking country	-0.29**	-0.17	-0.42***
	(0.10)	(0.09)	(0.11)
Foreign TA from English speaking country	-0.25	-0.05	-0.22*
	(0.13)	(0.18)	(0.11)
F-test for equality of coefficients	0.77	0.50	0.07
Mean dep. var.	3.97	4.102	4.139
SD dep. var.	.59	.57	.50
Observations	761	822	300
Class environment			
Foreign TA from non-English speaking country	-0.29**	-0.24**	-0.29***
	(0.09)	(0.08)	(0.08)
Foreign TA from English speaking country	-0.24	-0.07	-0.16
	(0.14)	(0.16)	(0.09)
F-test for equality of coefficients	0.73	0.24	0.22
Mean dep. var.	4.32	4.20	4.35
SD dep. var.	.44	.55	.37
Observations	761	822	300
Undergraduate student learning			
Foreign TA from non-English speaking country	-0.03	-0.04	-0.22***
	(0.14)	(0.09)	(0.06)
Foreign TA from English speaking country	-0.27	0.01	-0.10
	(0.16)	(0.11)	(0.06)
F-test for equality of coefficients	0.16	0.66	0.07
Mean dep. var.	3.98	3.98	3.94
SD dep. var.	.46	.54	.36
Observations	450	580	300

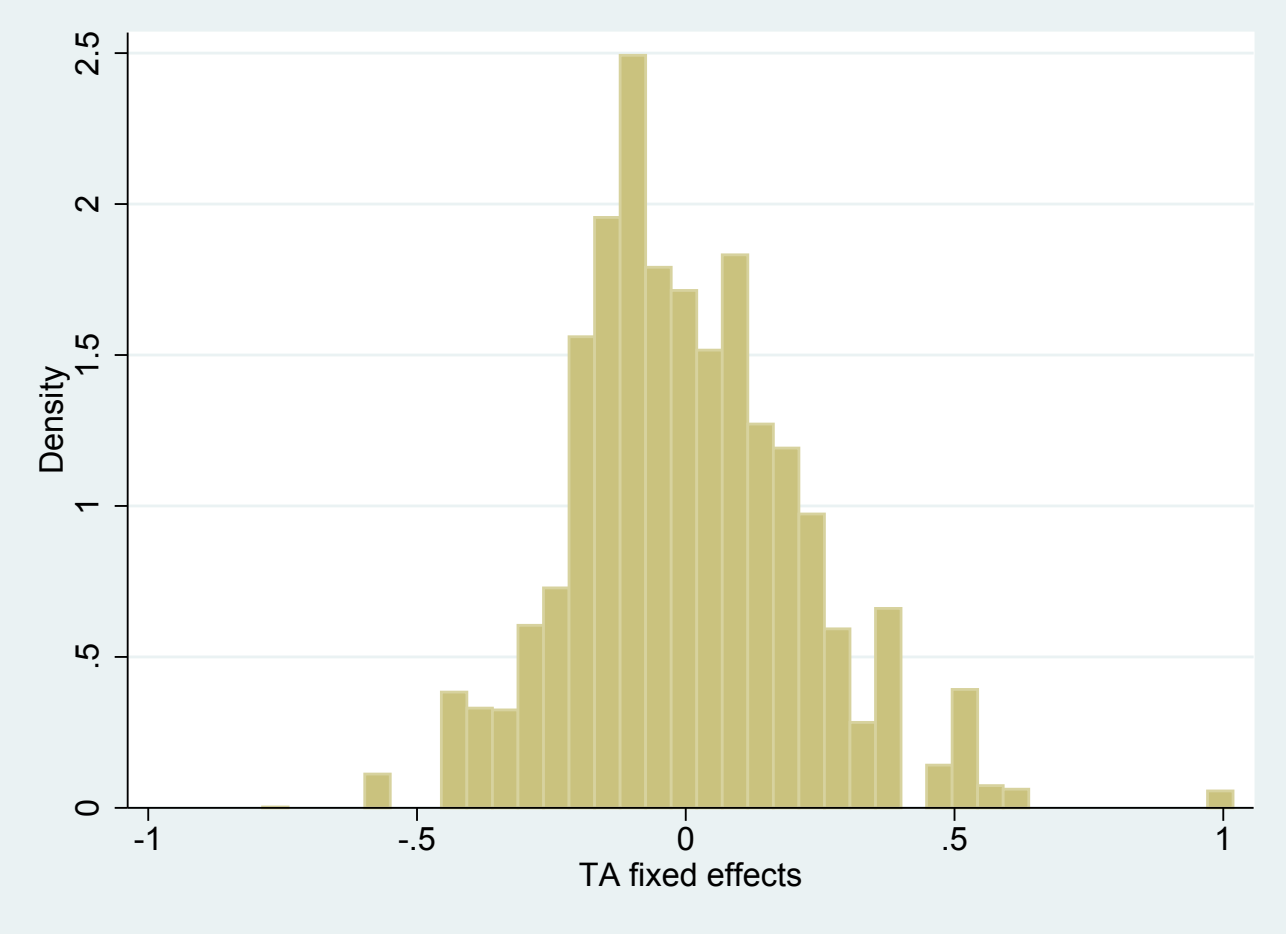
Notes: All specifications control for TA gender, race, age, times taught before, average undergraduate student characteristics, section time, and day of section. Course and term fixed effects are included and the standard errors are clustered by TA.

Table 6: Results for undergraduate student outcomes (OLS regression models)

VARIABLES	(1) Discussion	(2) Laboratory	(3) Full course
Grade			
Foreign TA from non-English speaking country	-0.04	-0.03	-0.03
	(0.02)	(0.02)	(0.03)
Foreign TA from English speaking country	-0.05	-0.04	-0.04
	(0.03)	(0.03)	(0.05)
F-test for equality of coefficients	0.69	0.77	0.88
Mean dep. var.	2.80	3.09	2.59
SD dep. var.	0.90	0.82	0.99
Observations	21,800	19,889	8,285
Ever declare STEM major			
Foreign TA from non-English speaking country	0.03*	0.00	-0.01
	(0.01)	(0.01)	(0.01)
Foreign TA from English speaking country	0.00	-0.00	-0.00
	(0.02)	(0.01)	(0.03)
F-test for equality of coefficients	0.09	0.57	0.70
Mean dep. var.	0.53	0.69	0.51
Observations	21,800	19,889	8,285
Ever graduate with STEM degree			
Foreign TA from non-English speaking country	0.00	-0.01	0.01
	(0.01)	(0.01)	(0.01)
Foreign TA from English speaking country	-0.00	-0.01	0.01
	(0.02)	(0.02)	(0.03)
F-test for equality of coefficients	0.70	0.98	0.90
Mean dep. var.	0.42	0.55	0.39
Observations	21,800	19,889	8,285
Unique undergraduate students	15256	13957	7729

Notes: All specifications control for TA gender, race, age, times taught before, undergraduate student characteristics, section time, and day of section. Course and term fixed effects are included and the standard errors are two-way clustered (undergraduate student and TA level).

Figure 3: Distribution of TA fixed effects



Notes: The distribution of TA fixed effects is the variance of θ from Equation 3.

Table 7: Main course grade outcome

TA effect	
SD(Calc I)	0.145
	(0.021)
SD(Calc II)	0.133
	(0.020)
Corr(Calc I, Calc II)	0.758
	(0.185)
Section effect	
SD(Calc I)	0.165
	(0.021)
SD(Calc II)	0.114
	(0.026)
Corr(Calc I, Calc II)	0.420
	(0.224)
Observations	694

Notes: Random effects models are estimated on section-level residuals. First stage models include TA and term fixed effects, in addition to individual controls and section average controls. Residuals are taken with respect to all variables other than TA fixed effects. Robust standard errors clustered by TA in parenthesis.

B Data Appendix

Table 8: Evaluation items for overall quality category

Overall, the instructor was an excellent teacher.
Overall, the TA was an excellent teacher.
Overall, the lab instructor was an excellent teacher.

Table 9: Evaluation items for TA effort category

The exams were returned in a reasonable amount of time.
Graded assignments (e.g. exams, papers) were returned in a reasonable amount of time.
The instructor was accessible to students outside of class.
The instructor handled questions well.
The TA handled questions well.
The instructor was open to contributions from all class members.
The lab instructor used techniques to foster class participation.
The instructor seemed well prepared for each class.
The instructor was well-prepared for each class.
The TA seemed well prepared for each class.
The instructor explained material clearly and understandably.
The instructor gave clear explanations.
The instructor presented material clearly in lectures/discussions.
The instructor delivered clear, organized explanations.
The TA gave clear and understandable explanations.
The lab instructor gave clear explanations.
The instructor used class time well.
The lab instructor used class time well.
The instructor helped me to understand the subject matter.
The instructor thoroughly understood the subject matter.
The instructor appeared to have a thorough knowledge of the subject.
The TA appeared to have a thorough knowledge of the subject.

Table 10: Evaluation items for environment category

Students felt comfortable asking questions.
The instructor treated students with respect.
Grades were assigned fairly and impartially.
Grading was a fair assessment of my performance in this course.
The TA graded papers (exams, homework) fairly.
The instructor was concerned that we learn.
The instructor was willing to help students outside of class.
The instructor gave individual attention to students in the class.
The instructor was sensitive to student difficulty with course work.
The instructor motivated me to work hard.
The instructor set high standards for students.
The instructor made the course difficult enough to be stimulating.
The class meetings were stimulating and informative.
This course increased my desire to learn more about this subject in the future.
I can see myself furthering my education in this area.
I deepened my interest in the subject matter of this course.
I developed enthusiasm about the course material.
The instructor was accessible to students outside of class.
The instructor had regular office hours and was available at those hours.
The instructor was willing to help students outside of class.
The instructor suggested specific ways students could improve.
The instructor kept students informed of their progress.
The instructor told students when they had done especially well.
The instructor made the course interesting.
The instructor seemed to enjoy teaching.
The instructor was enthusiastic.
The instructor maintained an atmosphere of good feeling in class.
I was very satisfied with the educational experience this instructor provided.
I would take another course with this instructor.
The instructor was enthusiastic about the subject matter.
The instructor was friendly.
My teacher demonstrates a strong commitment to teaching.
My teacher is fair and impartial when dealing with me.
The instructor was confident and in control of the class.
Students' difficulty with the material was recognized.
The instructor showed a genuine concern for the students.
The instructor knows me by name.
The instructor suggested specific ways students could improve.
The instructor was skillful in observing student reactions.
The lab instructor kept students informed of their progress.
The lab instructor set high standards for students.
The lab instructor taught in a manner that served my needs as a student.
The instructor brought out the best in me as a student.
The instructor encouraged student participation in an equitable way.
The instructor made good use of examples and illustrations.
The instructor made me feel known as an individual in this course.
The instructor made the course interesting.
The instructor maintained an atmosphere of good feeling in class.
The instructor responded effectively to student difficulty in class.

Table 11: Evaluation items for undergraduate student learning category

I learned a great deal from this course.
I gained a good understanding of concepts/principles in this field.
I deepened my interest in the subject matter of this course.
I developed the ability to communicate clearly about this subject.
I learned to apply principles from this course to new situations.
I learned a great deal in this laboratory.
I learned a great amount of substantive material.
I learned a great deal from this course.
I learned a great deal in this laboratory.
I learned to apply principles from this course to new situations.
I gained a good understanding of concepts/principles in this field.
I gained valuable experience working in teams in this course.
I increased my ability to analyze and interpret data.
I increased my ability to apply math and science knowledge to engineering problems.
I increased my ability to collect original data.
I increased my ability to design and conduct experiments.
I increased my ability to formulate, and solve engineering problems.
My confidence in my design abilities increased because of this course.
My oral communication skills improved because of this course.
My writing improved because of this course.
Course improved my ability to communicate technical information, designs, and analyses.

C Introductory STEM courses

C.1 Mathematics

The university considered offers four Math sequences.[34] My analysis only consists of the courses that are part of the standard sequence since the other courses are taught by entirely by faculty members. The standard sequence is taken by undergraduate students who plan to major in sciences or engineering and it contains the following courses: Calculus I, Calculus II and Calculus III. Calculus I and II consist of only lectures, while Calculus III has both a lecture and a laboratory (see Table 12). Out of the instructors teaching Calculus I, 64.3 percent are graduate students, while 68.5 out of the instructors teaching Calculus II are graduate students. The rest of the instructors are a combination of lecturers, post doctoral students and non-tenure track faculty. Only approximately

[34]The Math sequences offered are: the standard Math sequence, the applied honors Calculus sequence, the honors Calculus sequence and the honors seminar Math sequence.

2 percent of the instructors are tenure track faculty. As for Calculus III, 99.17 percent of the laboratories are taught by TAs while the lectures are taught by other types of instructors (professors, lecturers, etc.). All of the three courses considered have uniform exam dates. The exams are not multiple choice, but the TAs grade the exams together using the same answer key.

Table 12: Mathematics Courses Considered

Course	Components
Calculus I	LEC (taught by TA)
Calculus II	LEC (taught by TA)
Calculus III	LEC + LAB

C.2 Physics

The Physics Department offers three introductory course sequences. The Fundamental Concepts of Physics Sequence is comprised of the following courses: General Physics I, Elementary Lab I, General Physics II, Elementary Lab II, Waves Heat Light, Waves, Heat and Light Lab(see Table 13). General Physics I covers classical mechanics, while General Physics II covers electricity, magnetism, optics, and introduces concepts in modern physics. This sequence is designed for prospective physical science and engineering undergraduate students.[35] All the Physics laboratories have grades separate from the courses they pertain to. The exams for introductory classes take place at the same time. Most of the introductory courses have three midterms and a final exam.

General Physics I and II both consist of a lecture and a discussion session. Since only 4.46 of the discussion sessions in General Physics I are taught by TAs and none of the ones for General Physics II is taught by TAs (they are taught by a combination of professors (full, assistant or associate) and lecturers), I do not consider these two course for my analysis. Elementary Lab I is taught by 84.97 graduate students and it is a two-hour weekly laboratory designed to accompany General Physics I. Elementary Lab II contains a two-hour weekly laboratory that is taken at the same time with General Physics II. 84.69 of the instructors for the Elementary Lab II are TAs. The grade of this course is based on class performance and laboratory reports submitted each lab session (10 lab experiments in total). The lab courses have multiple choice quizzes graded by each TA. In addition to these quizzes, each section also has laboratory worksheets that are graded on an answer key made by the TAs.

[35]The university also offers a sequence for prospective life sciences students and one for honors students.

Table 13: Physics Courses Considered

Course	Components
Fundamental Concepts of Physics Sequence	
General Physics I	LEC+ DISC
Elementary Lab I	LAB
General Physics II	LEC+ DISC
Elementary Lab II	LAB

C.3 Chemistry

The general sequence (see Table 14) for undergraduate students interested in the sciences, engineering or medicine starts with either General Chemistry or Structure and Reactivity I, depending on how strong their Chemistry background is[36]. General Chemistry has a discussion and a lecture, where 97.68 percent of the discussions are taught by TAs. The General Chemistry Lab I consists of a discussion session and a laboratory, both taught mostly by TAs (95.82 percent and 95.26 respectively of TA-led sections). Structure and Reactivity I contains a discussion session (96.13 percent of discussions are taught by TAs) and a lecture. The course Investigations in Chemistry is made up of a laboratory and a lecture. Out of all the laboratory sessions, 78.09 percent are taught by TAs. The Synthesis and Characterization of Organic Compounds is composed of a lecture and a laboratory (85 percent taught by TAs). Structure and Reactivity II contains a lecture, laboratory and discussion session. The laboratory is taught in proportion of 84.35 by TAs and all discussion sessions are taught by TAs. There are no exams for the lab courses and the laboratory reports are graded by each TA using an answer key. The exams for the General Chemistry are multiple choice and scantron-graded, while the exams for Structure and Reactivity II are not multiple choice and grading is done together by all the TAs teaching the course, using a grading system set by the professor teaching the lecture.

C.4 Biology

Undergraduate students interested in majoring in biological sciences take the Introductory Biology Sequence (see Table 15). The first Biology course, Ecology/Evolution and Molecular, contains a lecture and a discussion session and 78.88 percent of the discussion sessions are led by TAs. The second Biology course, Introductory Biology - Molecular, Cellular, and Developmental, contains a lecture and a laboratory. The majority of the discussion sessions are taught by TAs (71.96 per-

[36] Students who took Chemistry AP credits in high school are advised to start with Structure and Reactivity I.

Table 14: Chemistry Courses Considered

Course	Components
General Chemistry	LEC+DISC
General Chemistry Lab I	LAB
Structure and Reactivity I	LEC+DISC
Investigations in Chemistry: Laboratory	LEC+LAB
Structure and Reactivity II	LEC
The Synthesis and Characterization of Organic Compounds	LAB+LEC

cent). These two courses are supplemented by an Introductory Biology Lab, with is taught by TAs almost entirely (94.97 of them are TA-led). Both of the Introductory Biology courses (Ecology and Evolution/ Molecular, Cellular, and Developmental) have multiple choice, scantron-grade exams (with the possibility of some short answers as well). The Introductory Biology laboratory contains two quizzes graded by each TA using an answer key provided by the lecture instructor. The older Biology introductory course contains both a discussion and a laboratory, both taught by the same TA, with 93.37 of labs/discussion sessions led by TAs.

Table 15: Biology Courses Considered

Course	Components
Introductory Biology Sequence	
Introductory Biology- Ecology and Evolution	LEC+DISC
Introductory Biology - Molecular, Cellular, and Developmental	LEC+DISC
Introductory Molecular Bio-Engineering	LEC
Introductory Biology Lab	LAB
Older courses	
Introductory Biology	LEC+DISC+LAB
Honors Introductory Biology	LEC+DISC
Introductory Microbiology	LEC+LAB

C.5 Engineering

All undergraduate students planning to major in Engineering are required to take a multitude of courses in different fields, including Mathematics, Physics, Chemistry and Engineering. Two of the required courses in any first year Engineering program are Introduction to Engineering and Introduction to Computing and Programming (see Table 16). While Introduction to Engineering contains both a discussion session and a laboratory, only $25, 58$ percent of the labs and only 4.87

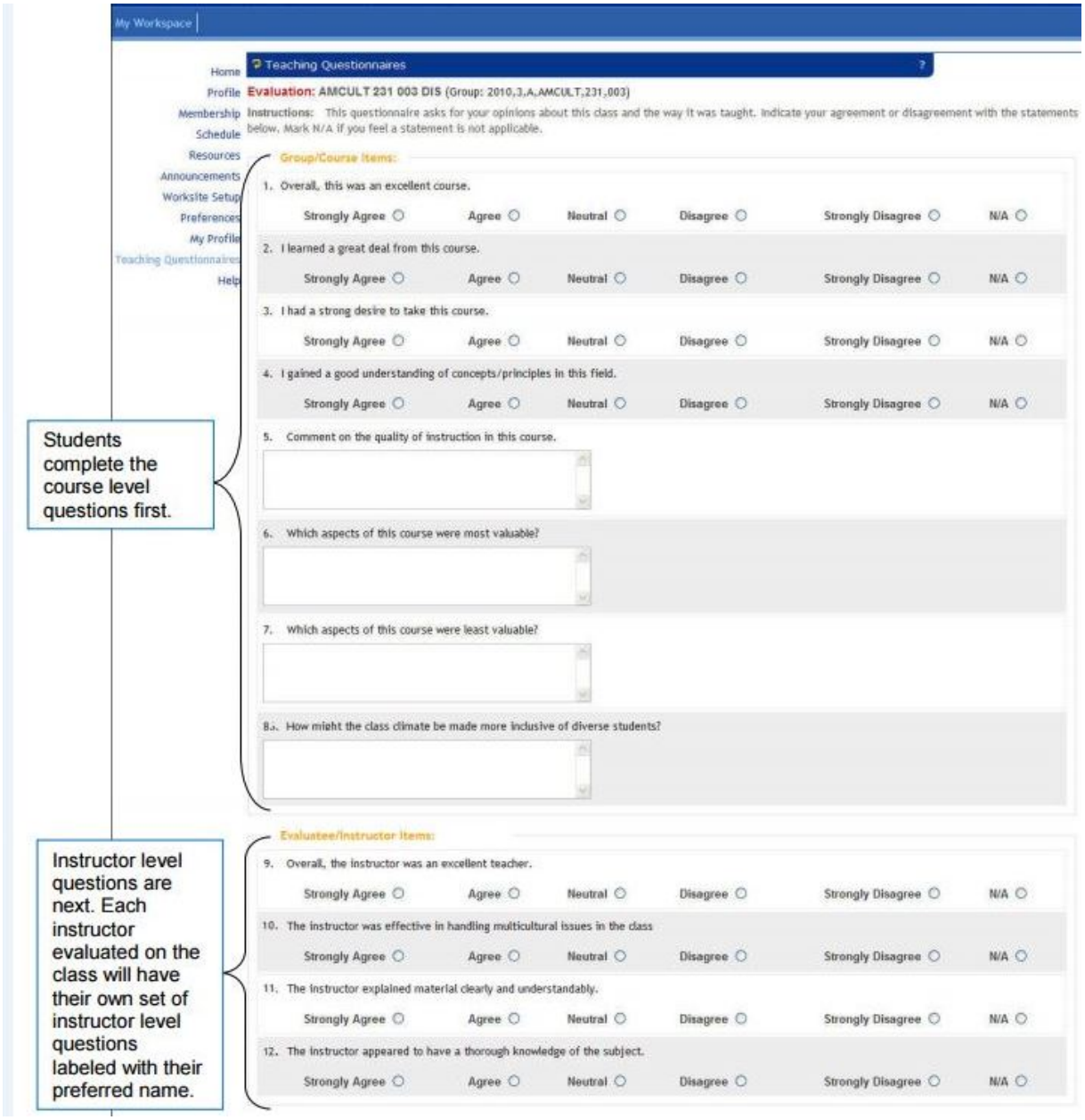
percent of the discussions sessions are taught by TAs. Therefore, I disregard this course and only consider Introduction to Computing and Programming, where TAs led 90.36 of the laboratories. Another reason for not including the Introduction to Engineering course is that the course does not have exams, but rather team projects, making the course too different than all the other courses considered in my analysis. The Introduction to Computing and Programming course has exams that are a combination of multiple choice questions and shorts answers and are graded by the TAs and the professors.

Table 16: Engineering Courses Considered

Course	Components
Introduction to Engineering	LEC+DISC+LAB
Introduction to Computing and Programming	LEC+LAB

D Student evaluation of teaching questions

Figure 4: Student evaluation of teaching questionnaire



E Computational example for the calculation of the median evaluation score

This section illustrates the calculation of the median score for each evaluation question based on the teaching evaluations filled out by the undergraduate students in each course.

Table 17: Example of student evaluation scores

Score	1	2	3	4	5
f	3	8	2	5	1
cf	3	11	13	18	19

Table 17 shows the student evaluation answers for a hypothetical question, the frequency and cumulative frequency of these answers. The median is defined as the point where or below where exactly 50 percent of the cases fall [Hays, 1973]. This implies that the frequency at the median should be exactly half of the total number of observations. Based on this, the median would divide the distribution into halves, with 19/2 scores above and 19/2 scores below the median. The scores don't quite divide themselves into two groups, and as seen above the median would fall somewhere in the interval containing 2. The upper and lower limits of this interval are 1.5 and 2.5, respectively. The median calculation is determined by interpolation by using the following formula:

$$m = L + c\frac{\frac{N}{2} - F_m}{b} \tag{7}$$

In the above formula, m =median, L =lower limit of the interval containing the median, c =the width of the interval containing the median=upper real limit−lower real limit, N =total number of responses, F = cumulative frequency b =number of observations within the interval containing the median. This implies:

$$\text{Median} = 1.5 + 1 * \frac{\frac{19}{2} - 3}{8} = 2.31 \tag{8}$$

Table 18: Sensitivity to the inclusion of different controls

Median evaluation scores			
Discussion sessions			
Foreign TA from non-English speaking country	-0.36**	-0.36**	-0.49***
	(0.13)	(0.12)	(0.13)
Foreign TA from English speaking country	-0.35	-0.31	-0.36
	(0.19)	(0.21)	(0.21)
Laboratories			
Foreign TA from non-English speaking country	-0.24*	-0.21*	-0.33**
	(0.10)	(0.10)	(0.10)
Foreign TA from English speaking country	-0.05	-0.06	-0.09
	(0.21)	(0.21)	(0.21)
Full courses			
Foreign TA from non-English speaking country	-0.52***	-0.58***	-0.48**
	(0.16)	(0.17)	(0.14)
Foreign TA from English speaking country	-0.27	-0.27	-0.30*
	(0.15)	(0.17)	(0.15)
Undergraduate student controls	Yes	No	Yes
TA controls	Yes	Yes	No
Grades			
Discussion sessions			
Foreign TA from non-English speaking country	-0.04	-0.01	0.01
	(0.02)	(0.03)	(0.02)
Foreign TA from English speaking country	-0.05	-0.12	-0.00
	(0.03)	(0.06)	(0.03)
Laboratories			
Foreign TA from non-English speaking country	-0.03	-0.04	0.01
	(0.02)	(0.02)	(0.01)
Foreign TA from English speaking country	-0.04	-0.05	0.00
	(0.03)	(0.03)	(0.03)
Full courses			
Foreign TA from non-English speaking country	-0.03	-0.06	-0.03
	(0.03)	(0.04)	(0.03)
Foreign TA from non-English speaking country	-0.04	-0.08	-0.04
	(0.05)	(0.05)	(0.05)

Notes: All specifications control for TA gender, race, age, times taught before, section time, and day of section. Course and term fixed effects are included. The median evaluation scores regressions also control for average undergraduate student characteristics and have the the standard errors are clustered by TA. The grade regressions control for undergraduate student characteristics and have the the standard errors two-way clustered (undergraduate student and TA level). Standard errors in parentheses.