This version: June 2015

Necessary Criticism? Peer Feedback and Teaching Performance

Carolina Mejía-Mantilla

Gabriela Rubio[1]

Abstract

We design, implement and evaluate a peer feedback program among teaching assistants (TAs) of a large public university with the objective of boosting teaching performance. The program was implemented during Fall 2012, and the measure of performance was the undergraduates' teaching evaluations of the TAs. We find a significant positive increase of one-half of a standard deviation in the TAs' evaluations for Winter 2013, and a smaller effect of one-fourth of a standard deviation for Spring 2013, showing that the program was successful but raising concerns on the medium/long run persistence of results. A detailed analysis of the individual components of the evaluations suggests that the intervention improved the TAs' communication skills, perceived concern, organization, scope and interaction. As an additional novel contribution, this paper also shows that undergraduates' satisfaction with the course also increased by almost one-half of a standard deviation for Winter 2013 and by one-fifth of a standard deviation for Spring 2013.

Keywords: Experiment, teaching skills, student evaluation, peer feedback.

JEL classification: I2, I23.

[1]cmejiamantilla@worldbank.org, grubio4@ucmerced.edu. We are grateful to Adriana Lleras-Muney, Kathleen McGarry, Day Manoli and Alex Whalley for their guidance and comments. We are also particularly grateful to Joanne Valli-Meredith, Director of the Evaluation and Educational Assessment Office, for her time and support since the beginning of the project. Special thanks to Sarah Reber and Paola Guliano for their insights on the experimental design. We are really grateful to the Department for the financial, logistic and general support, in particular to Joe Ostroy, Ed McDevitt (who was immensely patient with our requests), Janette Briceno-Ferrier, Nancy Blumstein, Jessica Perez and Ciril Bosch-Rosa. Alma Lopez provided excellent administrative help with the intervention. Comments from participants of the applied micro group meetings are also acknowledged. All remaining errors are ours.

1 Introduction

Graduate teaching assistants (TAs) are widely used in the American higher education system. Since 2000, they have steadily represented around 4.5% of the total employment of Universities, Colleges and Professional Schools, and slightly more than 50% of the total professional staff in these institutions (BLS 2002 and 2014). Around 70% of graduate students have had some teaching responsibility and, between 2001 and 2011, the number of graduate teaching assistants increased by 36%, compared to an increase of 19% of full-time faculty (NCES 2000, 2001, 2013).[2]

In spite of the widespread use of TAs, little research attention has been paid to the most effective practices for improving their quality of teaching.[3,4] Peer observation, defined as a "collaborative activity in which professionals offer mutual support by observing each other teach; explaining and discussing what was observed; sharing ideas about teaching, among others," is one potentially high impact intervention (Gosling 2005, Bell and Mladenovic 2008).

Despite the potential impact of peer evaluation there is a central measurement challenge in evaluating whether peer evaluation is *cost effective*. These programs have (mostly) been implemented using volunteer instructors in selected Universities, raising concerns regarding who is volunteering and what other changes are introduced in these schools. Are these the most enthusiastic instructors? Are these lagging behind in performance? The adoption of this practice might also be bundled with other policies introduced at the same time that also affect student outcomes; for instance, instructors might improve their teaching performance if they believe that it will impact their tenure evaluation (Thomas et al. 2014). This paper addresses these concerns by designing and implementing a randomized controlled trial (RCT) to establish a causal relationship between peer feedback and teaching performance of TAs. We address the selection problem by randomly assigning all the TAs, in one Department of a large public university, to either a control group or to a treatment group - members of the latter observed and provided feedback to each other. And importantly, during this period, the Department did not introduce any other program aiming to improve teaching practices

[2] Although there has been limited attention given to teaching assistants (TAs) in the literature, there is evidence that TAs have a positive effect on students' performance (Hanushek 2007, Borjas 2000, and Watts and Lynch 1989).

[3] About half of the economics departments that permit graduate students to teach their own courses require no formal departmental training in teaching (McCoy and Milkman 2010; Walstad and Becker 2010) and almost half of their graduate students report that they received no preparation for teaching (McGoldrick, Hoyt, and Colander 2010).

[4] In recent years there has been an increasing interest on measuring the value added by universities to their students, while recognizing that this is a challenging task (Cuhna and Miller, 2014) and that some of the student-level measures of institutional performance are affected by several factors, among them non-instructional characteristics of the institutions - academic support, student services and research (Webber and Ehrenberg, 2010). TAs should be considered an invaluable resource in some of these areas, by having a closer interaction with students than instructors (in larger universities class size might reach the triple digits, while discussion section are typically limited to 30 or less students per section) and by potentially being perceive as more approachable by students (TAs tend to be younger than professors).

of TAs.

To date work exploring interventions for improving teaching practices at the higher education level have mainly been qualitative; thus whether peer evaluation has a casual effect on student learning and is cost effective remains an open question. Peer observation provides competent assessment from colleagues who perform the same activity, possess comparable academic qualifications, and are familiar with the context in which the teaching is taking place. In addition, it is a practice that can be relatively inexpensive and easy to implement within any Department, making it an attractive and *low cost* intervention.[5]

The intervention targeted 55 TAs during the Fall Quarter of 2012. Of these TAs, 78% of the (32) TAs assigned to the treatment group agreed to participate. The treatment consisted of three elements: i) a two hour workshop taking place during the first week of Fall 2012, covering the key aspects for providing constructive criticism; ii) twice during this quarter, each TA in the treatment group was observed while teaching and received detailed written feedback by two TAs in the same group;[6] iii) in turn, also twice during the same quarter, each (treated) TA observed and gave feedback to two other (treated) TAs while teaching. To promote participation and to justify the hours of additional work, the TAs in the treatment group were compensated with a cash reward of $100 each, contingent on complying with all the activities of the RCT.

The effect of the program was studied using data from the undergraduate students' evaluation of the TAs for Fall 2012, Winter 2013 and Spring 2013.[7] The results from Fall 2012 indicate the TAs did not have enough time to adjust their teaching practices during the quarter of the intervention; the effect is positive but non-statistically significant.[8] The intervention, however, had a large and statistically significant effect on the teaching evaluations of Winter 2013 - we estimate an increase of one-half of a standard deviation in the TAs' evaluations; and it had a positive, although imprecisely estimated, effect for Spring 2013 - of one-quarter of a standard deviation. We also evaluated the effect on the specific skills rated by the undergraduate students as part of the overall assessment of the TA,

[5]Throughout this paper we use the terms peer observation, peer feedback and peer review interchangeably.

[6]We created two formats ("observation" form and "feedback" form) in order to help providing easy, structured and constructive feedback. Details on the implementation of the RCT are presented in section 2.

[7]The literature has criticized the use of student evaluations of teaching as a way of assessing the teaching quality of faculty (Becker and Watts 1999, Siegfried and Walstad 1998, McPherson, Jewell, and Kim 2009; Kherfi 2011; Braga, Paccagnella and Pellizari, 2014; Beleche, Fairris and Marks, 2012; among others). The evaluations might be influenced by the faulty members and might respond to characteristics not related to teaching (e.g., race, gender, age) or they might be negatively correlated with teaching effectiveness. In our case, the characteristics of TAs are balance across groups, limiting the possibility that they drive any estimated effect. In addition, TAs are not in charge of the grading scale or of determining the level of difficulty of assignments - mitigating the concern that they can manipulate their evaluations - and professors usually determine the scope and material covered in TA sections, which also mitigates the concern that students penalize TAs assigned to more (stricter) effective instructors - moreover, our results hold when controlling for course (instructor) fixed effects.

[8]The peer review program took place between weeks four and nine of the ten-week quarter, leaving only a few sections for the TAs to internalize the feedback and adjust their teaching practices, more details on the timeline of the intervention are found in sections 2 and 3.

finding that the intervention had a positive effect over the dimensions of what is considered good teaching: concern about the students' learning, organization of the class, interaction with students, and communication skills. Similar to previous our results, the effects on each category are particularly large for Winter, and, although there is persistence for Spring, the magnitude is smaller and not-statistically significant.

Do TA teaching practices affect overall satisfaction with the course? To see, we perform one novel (and complementary) analysis by studying the effect on average course evaluation.[9] Similar to our previous results, the program had not effect during Fall 2012, but it increased the Winter 2013 evaluations by almost one-half of a standard deviation and the Spring 2013 evaluations by one-fifth of a standard deviation. These results are novel by showing that improving TAs performance might improve overall students' satisfaction with the courses.[10] Finally, we conducted a post-intervention qualitative survey that suggested that the TAs considered that the program was valuable, although they indicated that learning also happened by observing their peers teach.

This paper contributes to the literature studying the behavior and impact of TAs (Park, 2004). Tuckman (1975) explores the impact of TAs on students' outcomes, finding that TAs are as effective as experienced faculty. Watts and Lynch (1989) study several factors affecting students' achievement; their findings suggest a negative effect of non-native English speaker TAs on students' output. In contrast, Norris (1991) finds that non-native English-speaker TAs outperformed natives after controlling for "teaching experience." Finally, Borjas (2000) conducts a similar study investigating the impact of foreign-born TAs on students' grades, finding that limited English proficiency adversely affects students' grades. Nonetheless, he finds no such performance gap for those students who believe that foreign born TAs are better prepared than native born TAs. These results suggest two key findings: (i) TAs' quality matter for students' performance in a given course; (ii) increased effort, better preparation, or increased teaching skills seem to be as important as other characteristics of the TAs compensating, for instance, for a lower proficiency in English. Our study contributes to this literature by showing a communication improving intervention can affect postively affect teaching qualit by all TAs, not just non-native english speakers

Our paper also contributes to a growing literature assessing the effects of monetary and

[9]This measure is calculated using the grade that undergraduate students assign to the course in general (as opposed to the discussion section or the TA). It might be considered a weighted average of student satisfaction with the course, the professor and the TA.

[10]The determinants of student satisfaction with their courses, instructors and universities are subject of great deal of interest among scholars because they might determine student demand, the quality of students (and thus, competition for talented students), and placement, among many others (Lenton, 2015; Numberg, Schapiro and Zimmerman, 2012; Cheng and Marsh, 2010). Scholars have shown that higher expenditure on academic staff (per student) is not necessarily associated with higher student satisfaction, one possible reason is that it not necessarily signal higher quality of teaching (Lenton, 2015). Our results show a potentially cheap and easy intervention for boosting student satisfaction (the program might be easily implemented by modifying the contracts of TAs).

non-monetary incentives on teaching practices.[11] The education literature has explored the effectiveness of alternative interventions aimed at improving teaching practices.[12] Gosling (2005), highly influential in this area, identifies three main types of peer review of teaching: (i) evaluative model, which refers to senior staff or administrative evaluators assessing the performance of junior members, mostly used for promotion or to confirm tenure; (ii) developmental model, involving the use of expert educational developers with the objective of improving teaching competencies during initial (or reinforcing) training of staff members; and (iii) collaborative, colleagues of similar seniority observing each other with the goal of improving teaching practices through dialogue, and self and mutual reflection, it might be implemented in an ongoing basis and mutually benefit both parties involved (reviewer and reviewee). Our design falls into this last classification.

Earlier studies focused on understanding the effect of peer training on teachers and instructors (Gilbert and McArthur 1975, Caroll 1980, Dalgaard 1982). More recent attempts have focused on studying peer review programs implemented in selected Universities of Australia, the U.K. and the USA. The most common methodology of these studies has been to engage small groups of instructors or lecturers in peer review exercises and assess their success mainly using interviews of the participants, which inquire about satisfaction with the exercise, concerns about the program, and usefulness, among others.[13] Supporters argue that because it is based on constructive feedback and monitoring among colleagues, these interventions result in improvements in teaching and the enhancement of confidence (Bell 2005), development of collegiality (Quinlan and Akerlind 2000), and integration into the department (Allen 2002). Despite the overall positive appraisal of peer feedback, some potential pitfalls have been identified. It may be considered intrusive and uncomfortable, include a subjective component which might bias the assessment and it might be challenging to engage in critical reflection, and to provide and to accept feedback. (Allen 2002, David and Macayan 2010, Bell 2005).

[11] This literature has consistently found that better teaching practices (higher teacher quality) have a positive effect on student's outcomes (e.g., see Hanushek 1986, 1995; Ingersoll 2003; Rice 2003). Most of these results are found in primary and secondary education. Nonetheless, there is still a gap in understanding how to improve quality of teaching, specially in tertiary education (Allgood, Walstand and Siegfried (2015) summarizes the main research findings about teaching economics to undergraduates).

[12] Some examples are training sessions, the assessment of teacher performance by a third party, peer observation by colleagues in the same discipline, and self-reviews, studied mostly from a qualitative perspective (Caroll 1980, Sparks 1986, Lawrenz et al. 1992, Croteau and Hoynes 1991, Robinson 2000, Gosling 2005, Gosling and O'Connor 2006 and 2009, Bell and Mladenovic 2008, Sachs and Parsell 2014)

[13] Thomas et al. (2014) summarizes a total of 27 papers found in the education literature, a few examples are: Attwood, Taylor and Hutchings (2000) implemented a program in 1996 using volunteer Chemistry faculty members from 7 USA universities; Lomas and Nicholls (2005) examined the introduction of peer review of teaching in a U.K. university through individual interviews with 100 faculty members; Kell and Annetts (2009) assessed the perceptions about these programs using a group of 20 faulty members within the Department of Physiotherapy at Cardiff University, finding that newer faculty members tended to perceive it as audit-like, while senior faculty tend to consider it as beneficial for personal and professional development; Hammersley-Fletcher and Orsmond (2004) evaluate two systems of peer observation implemented in a British university using academics from the Law Faculty and the School of Science. They interviewed a group of 9 volunteer faculty members who showed concern about negative feedback and criticism.

The rest of the paper is organized as follows. The details of the sample and the experimental design are presented in Section 2, while the descriptive statistics and the results are discussed in Section 3. Section 4 contains the results on the complementary analyses and section 5 concludes.

2 Experimental Design

The intervention took place during the Fall Quarter of 2012 in the Economics Department of a large public university.[14] The class enrollment allowed for a total of 55 TAs to be eligible for the intervention. Every graduate student with a Teaching Assistant Fellowship is responsible for teaching two discussion sessions per week of a given course throughout the academic quarter. The eligible TAs were randomly selected into one of two groups: the control or the treatment group. Those in the treatment group had the option of declining to participate in the intervention. This section provides specific details of the experimental design and the recruitment process of the intervention.

The design and evaluation of the program was approved by the Office of Instructional Development, which is in charge of conducting the TAs' evaluations, of implementing and assessing changes to instructional practices, and of providing (voluntary) training workshops, among other tasks. We want to empathize that our experiment borrows from the literature in education trying to mitigate the main problems highlighted by these scholars: i) we use a structured and standardized form for the feedback; ii) we provide training trying to teach TAs how to give nonjudgmental suggestions; and iii) we maintain anonymity of observers and observees at the department level (in case the TAs fear punishment for giving or receiving a low score).

2.1 Experimental Design

Collaborative peer review programs are designed in such a way that within a school or department, teachers and/or professors evaluate each other. Along those lines, the TAs assigned to the treatment group acted both as observers and observed subjects. Within the treatment group, each TA was observed and evaluated by two other TAs - also belonging to the treatment group - while teaching a discussion section. Each TA was observed twice during the quarter and received a total of four feedback forms. The evaluation took place from weeks four to nine of the ten-week quarter. The dates were adjusted according to the midterm calendar of each course (to avoid the high rate of absenteeism in discussion sections immediately following a midterm, and to avoid a session where the TA merely reviews the midterm questions). In an attempt to prevent any special preparation by the TA being evaluated, the observation date was only announced to the observer and not to the observed TA; however, TAs could have (imperfectly) inferred the timing based on their

[14] Each quarter in composed of ten weeks for instruction and one week for final exams.

own observation dates.[15]

All observations/assessments followed a detailed format (presented in Figure 1) which emphasizes the factors related to teaching effectiveness considered to be under the control of the TAs. Observers were instructed not to interfere with the discussion session in any way and to submit written feedback to the project managers within the next couple of days. We designed an additional form (feedback form shown in Figure 2) similar to the observation form but conceived for including specific suggestions in each of the categories related to teaching effectiveness. The aim was to suggest the TAs concrete actions for improving their discussion sections - the feedback given to the TA was meant to have a constructive and useful tone, instead of being mere criticisms.

The sample size constrained the experimental design; in particular, all the observers belonged to the treatment group, therefore the intervention also involved observed TAs acting as observers of the discussion sections of two fellow TAs. This feature of the intervention means that they could have learnt or adopted teaching practices useful to them while teaching or preparing for class. This characteristic of the design must be emphasized, feedback was not the only component of the treatment - members of the treatment group were also exposed to the teaching practices of their peers, an experience which may also have had an impact on their teaching.

An additional concern was that most graduate students have no training or previous knowledge in assessing teaching performance, which might reduce the efficacy of the feedback.[16] In order to mitigate this problem, the Office of Instructional Development provided assistance designing and implementing a training workshop which took place before the observations started and was conducted by an education professional (who is an expert in evaluating teaching skills).[17] The contents of the training session, and of the observation and feedback forms were closely related to the basic components of good teaching skills according to the framework developed by Marsh (1983) and traditionally cited in the education literature. According to this framework, there are eight factors commonly related to teacher effectiveness: i) organization/clarity, ii) group interaction, iii) instructor enthusiasm, iv) learning/value, v) breadth of coverage, vi) examinations/grading, vii) assignments/readings, and viii) workload/difficulty. In the context of this study, only factors

[15]The experiment was designed to have the two observers attending the same session; however, in a few cases it was not possible due to scheduling conflicts, but we made an effort to at least make them observe during the same week.

[16]Salemi and Walstad (2010) implemented a training program known as the Teaching Innovations Program (TIP) in economics, which was conducted over a six-year period for 335 economics professors, funded by the National Science Foundation. In its first phase, economics instructors attended a workshop to learn how to use interactive teaching methods. The second phase consisted of a voluntary program of online instruction and mentoring to help the economics instructors use a teaching innovation in their classrooms. A third phase offered participants an opportunity to contribute to the scholarship of teaching economics by making conference presentations or writing papers about their pedagogical experiences. At the end of the study, it was revealed that 95% of the participants thought that the program improved their teaching.

[17]The trainer emphasized that both negative and positive aspects should be brought up and any criticism should be accompanied with a suggestion on how to improve.

i) through v) are relevant, since the remainder are not under the TAs' influence. After each observation round, the project manager emailed the two anonymous feedback forms to the observed TA, who was required to acknowledge their receipt.[18,19]

Finally, a potential problem was that the cash compensation would not be enough incentive for providing thoughtful and careful feedback. In order to promote better quality and more useful feedback, each treated TA was assigned two different observers who would observe her simultaneously and provide feedback about the same sessions - this increased the likelihood that a TA would receive useful feedback from at least one person, as well as providing some peer pressure for the evaluators to take the process seriously. In addition, we performed random spot checks to ensure that the evaluators were present (at the beginning of the intervention, we informed the evaluators about this component of the program).[20]

Lastly, the TAs in the treatment group received a cash reward of $100 as compensation for approximately 5 hours of work throughout the quarter (an effective rate slightly lower than the $25 hourly rate TAs generally receive for teaching). The cash compensation was conditional on participating in the initial training workshop, attending the assigned sections, and returning the observation/feedback forms on time.[21]

2.2 Recruitment Process

With the help of the Department's Graduate Advisor, we contacted via email all the TAs assigned into the treatment group and informed them that they had been selected to participate in a "teaching training program," for which they would be compensated if they agreed to participate. Of the initial 32 TAs assigned to the treatment group, 25 agreed to participate in the program, which translates into a take up rate of 78%. The TAs who agreed to participate attended the two hour workshop during which we explained the activities expected from them and the compensation scheme, they would receive a the cash compensation at the end of the quarter if they agreed to: i) observe and provide feedback to two fellow TAs twice during the quarter, using the forms previously described; and, ii) be observed and receive feedback from two fellow TAs twice throughout the quarter. We emphasized that even though this project was supported and funded by the Department, there was no penalty for not participating and no additional reward for doing so. We also assured them that the feedback provided or received would be analyzed only by the

[18] In the qualitative surveys TAs indicated that they indeed read the feedback shortly after the observation took place.

[19] It should be noted that since factors vi) to viii) are not under the TAs' influence, the results documented in this paper are not likely to be due to "soft-grading" (Kanagaretnam, Mathieu, and Thevaranjan 2003, Love and Kotchen 2010, among others), TAs do not determine the scale for grading. Moreover, professors usually have the same TA grading all assignments or exams of a given course in order to mitigate this kind of manipulation and to maintain consistency in grading across all sections.

[20] These components were added to mitigate low-quality feedback as a potential drawback of the intervention.

[21] 24 out of the 25 TAs participating in the treatment program complied with all the activities of the RCT. The TA who failed to attend all the sessions was excused from the very last observation due to health issues (this TA completed 3 out of 4 observations but received full payment).

research team and that the Department would only have access to the overall assessment of the program. This was done to mitigate any concerns about negative repercussions for individuals participating in the program, and to encourage honest feedback.[22,23]

3 Descriptive Statistics and Results

This section provides descriptive statistics of the sample of TAs that were eligible for the intervention, as well as the results of the intervention on the main outcomes of interest, the undergraduate students' evaluation of the TA. For Fall 2012, we had also access to the grades of the undergraduate students.[24] In addition, in section 4.1, we analyze the effect on a measure labeled "average course evaluation," which refers to the mean evaluation of the course calculated using only the subsample of students in a specific TA section. This measure might partially reflect student satisfaction with the course and the main instructor, which in turn might be affected by the performance of the TA.

The observations took place between weeks four and nine of the ten week quarter of Fall 2012. For the most part, all first round observation took place in weeks four and five (98%), but only 86% second round observations took place in weeks six or seven, the reminder of which took place in weeks eight and nine. The students' evaluations of TAs and professors usually take place during week nine or ten (last week) of the quarter; therefore, it might be the case that for those TAs who were observed later in the quarter there was not enough time to incorporate the second round of feedback before the evaluations of that same quarter took place.

We present the effects of the intervention on the Intent to Treat group (ITT) - all the TAs that were assigned to treatment and were offered the chance to participate in the program - and on the Treatment on the Treated group (ToT) - those TAs who agreed to participate in the program and actually received the treatment. The ITT group was selected at random, and is not subject to the concern that those choosing to participate in the treatment might be those who believe they will get a particularly strong benefit from the program and might, therefore, differ in unobserved ways from the TAs choosing not to participate.

3.1 Descriptive Statistics and Comparison of Means

In order to fully estimate the impact of the program, we analyze the effect of the program

[22]The groups were not set up for mutual observation (i.e. TA 1 observing TA 2, and TA 2 observing TA 1 in turn). Due to scheduling conflicts, in some cases we were not able to avoid mutual observation. However, we only disclosed the name of the observed TA to the two observers within a few days of each round. We also stressed that for the second round the observers could change. We implemented this measure for avoiding potential collusion among TAs (for instance, they could agree to give each other a high score and only positive reviews).

[23]After the workshop, all the attendees signed a consent form in which they agreed to be a part of the program and in which it was clear that failing any of requirements of the program would result in receiving no compensation at all.

[24]The measured used was the deviation of the TAs section grade average from the course average. For this outcome, the effect can only be identified in courses with many sections, mainly the introductory and lower division courses (see Table 2).

for three quarters: Fall 2012, Winter 2013 and Spring 2013. This allows us to study if the peer review program was successful at the time of the intervention, and one and two quarters after it took place. The allocation of TAships is made on a quarterly basis depending on the needs of the department and the availability of the graduate students. For the Winter 2013 Quarter, only four of the graduate students involved in the intervention did not have a TAship. Three of them belonged to the control group; the fourth was a non-complier from the treatment group. For the Spring 2013 Quarter, seven TAs (of the original sample) were not teaching. Three of them belonged to the control group and the other four were part of the treatment group - 2 non-compliers and 2 compliers. We perform a detailed analysis of attrition in section 4.2.

The TAs available for the intervention have a diverse background and the majority are not US citizens. Figures 3 and 4 show the country of origin and undergraduate major of the TAs in the sample. As illustrated in Figure 3, a large portion of the TAs, roughly 80%, come from outside the US, mainly from China, Korea and Latin America. Most of them (55%) majored in Economics for their undergraduate degree, or Economics and Mathematics (15%).[25] Figure 5 shows that the most popular field of concentration is Macroeconomics (40%), followed by Theory (24%), Labor (14%), Econometrics (11%) and Industrial Organization (11%).

Table 1 shows descriptive statistics on the covariates of interest. We focus on age, gender, being an English native speaker, PhD year, having a masters degree prior to starting the PhD, number of quarters taught at the university, having taught the same course in the past and having had a meeting with the TA coordinator. All these variables might affect the teaching skills of the TA and/or the perception of students. The last variable, "meeting with the TA coordinator", represents a prior effort of the Department for improving teaching skills: if a TA gets an average evaluation below seven, they meet with the coordinator who suggests possible strategies for improving their teaching style.[26] For Fall 2012, the table shows that the average age of the TAs is 27, approximately three fourths are male and, consistent with the information of country of origin, only 22% are native English speakers. In terms of teaching experience, they have taught an average of six quarters in the university, roughly half of them had taught the same course in the past and 14% had met with the TA coordinator.

The samples are very similar in the three quarters. The main differences are for the three variables capturing information on teaching experience. The average number of quarters taught increases from 6.4 in Fall 2012 to 7.12 in Winter 2013 and to 7.31 in Spring 2013. This change is the mechanic result of TAs teaching throughout the 2012-2013 academic year.

[25]The TAs are PhD students that are in their second to sixth year of the program. They are required to be making satisfactory progress in the program and not to be hired as research assistants or to obtain funding from other fellowships.

[26]The TA coordinator is typically a graduate student in her last year of the program who has shown exceptional teaching skills throughout the previous years.

The proportion of TAs teaching a course they have already taught in the past increases from 45% in Fall 2012 to 57% in Winter 2013 and to 63% in Spring 2013. This is mainly the result of TAs teaching a specific course for the first time during Fall 2012 being able to continue teaching the same course in subsequent quarters (introductory classes are offered every quarter). Finally, the proportion of TAs meeting with the TA coordinator increases from 14% in Fall 2012 to 34% in Winter 2013 and to 33% in Spring 2013.[27]

The assignment to the treatment and control groups was done randomly, but stratified by course when possible, as shown in Table 2. The table shows the distribution for the ITT group (where TAs were assigned to treatment but did not necessarily agree to participate) and the ToT group across the courses offered. The purpose of stratification was to minimize the effect of course-specific traits, such as difficulty, teaching skills of the main lecturer, or individual student interest in the subject, on the TA evaluations.

Table 3 explores whether our main covariates of interest are balanced across groups. The table presents the results for the ITT and the ToT groups, separately for each quarter studied. We find no statistically significant difference in means between groups (Table 3) for any of the eight variables, with the exception of "having a masters degree" in the Spring 2013 quarter (for both ITT and ToT).[28]

The main outcomes of interest, shown in table 4, are: (i) the student evaluations of the TA performance for Fall 2012, Winter and Spring 2013 - at the end of the quarter, undergraduate students fill out a TA evaluation form distributed by the Office of Instructional Development; and (ii) the Fall 2012 students' grades, more specifically, the deviation of the TA's section grade average from the average for the whole course (which in most cases comprises many TA sections) for Fall 2012.[29,30] Regarding the undergraduate student evaluations of the TA, we were particularly interested in the overall rating of the teaching assistant (mean TA evaluation) answered on a scale from 1 (Very Low) to 9 (Very High).[31] For Fall 2012, the average overall TA evaluation by section is 7.8. The average evaluations do not change for Winter 2013 and slightly improve for Spring 2013, increasing to 7.9.

Table 5 shows the difference in section average and median for the three quarters and

[27]For our regression analysis we use pre-treatment values for all covariates in order to mitigate potential concerns of multicollinearity: for instance, if treated TAs were less likely to meet with the TA coordinator as a result of the program, updating the indicator variable for meeting with the TA coordinator for subsequent quarters might lead to estimating a smaller (or no effect) of the program. All results hold if we update these covariates, these results are available upon request.

[28]Even though covariates are balanced between the treatment and the control groups, we explore whether there was selection into treatment. That is, whether the compliers - TAs assigned to the treatment group who agreed to participate in the intervention - are inherently different from the non-compliers - TAs those who were assigned to the treatment group but chose not to participate. The results are available upon request, but our analysis finds no evidence for selection into treatment in terms of the covariates.

[29]The purpose of using deviations from the course mean was to reduce the noise caused by differences between courses and to focus on differences between treated and non-treated TAs.

[30]Due to a few changes in personnel within the Department, we were not able to obtain the undergraduate students' grades for Winter 2013 and Spring 2013.

[31]Table 15 of the appendix shows a copy of the actual evaluations that the undergraduate students fill out.

the two groups, ITT and ToT. The tables also present the deviation of undergraduate students' grades from the course average grades for Fall 2012.[32] As shown in Table 5, for Fall 2012, peer feedback seems to have a positive treatment effect on the average evaluation of TAs, the magnitude is around 0.17, which is almost one-fifth of the standard deviation; nonetheless, it is not statistically significant. However, there is not any treatment effect on the undergraduate grades. These results (Fall 2012 Quarter) might not be surprising since the peer review program took place between weeks four and nine, leaving only a few sections for the TAs to internalize the feedback and adjust their teaching practices.

The results for Winter 2013 and Spring 2013 are larger in magnitude than those for Fall 2012, but only statistically significant for Winter 2013. We estimate an effect of 0.37 points for the Winter 2013 TAs' mean evaluation, which represent almost one-half of a standard deviation. For Spring 2013, the estimated effect is 0.24 - or one-fourth of a standard deviation in that quarter.

Comparable results are found for the ToT group. The effect during Fall 2012 over the average TAs' evaluation is around 0.16 points but still not statistically significant, and there is no effect on the final grades of students. The effect on the mean TAs' evaluations are 0.36 and 0.19 points for the Winter 2013 and Spring 2013 quarters, respectively - one-half of a standard deviation for Winter and one-fifth of a standard deviation for Spring (only statistically significant for Winter 2013).

The results of the ITT and the ToT for the Winter 2013 quarter are both statistically significant and large in magnitude. These findings suggest that the intervention was successful in the short run. Once the TAs had enough time to incorporate the suggestions made by their peers and to adopt the lessons from their own observations, they improved their performance. However, for Spring 2013, they are smaller in magnitude and are no longer statistically different from zero. These results raise questions about the persistence of such effects, indicating that follow up interventions may be required to sustain the short term results.

3.2 Regression Analysis

The effects of the intervention can be assessed by comparing outcomes across groups using an Ordinary Least Squares regression model. We focus on the average TA evaluation since it is the most direct assessment of the performance of teaching assistants. In this section, estimate the following specification:

$$y_{i,a} = \alpha + \theta Treat_a + \beta X_a + \delta_i + \varepsilon_{i,a} \tag{1}$$

where, $y_{i,a}$ is the outcome of interest for section i of TA a, $Treat$ is the Intent to Treat

[32] Note that the treatment is at the TA level and each TA teaches two sections.

(ITT) or the Treatment on the Treated (ToT) indicator at the TA level, X_a are a set of controls at the TA level, δ_i are course-specific fixed effects used in some specifications, and $\varepsilon_{i,a}$ are robust errors, clustered at the TA level. The coefficient of interest is θ, which should be an unbiased indicator of the causal effect of the intervention - the unobservable characteristics of the TAs should be distributed randomly across the groups due to the experimental design.[33]

Table 6 shows the ITT and ToT effects on the average TAs' evaluation and on the students' grades (more specifically the deviation of the TA average grade from the course average grade) of the Fall 2012 Quarter. Similar to the comparison of means of section 3.1, the estimated effect is positive but statistically insignificant. The empirical analysis suggests again that there was no effect of peer feedback on the students' performance; both ITT and ToT coefficients are very close to zero, and again not statistically significant.

It should be emphasized that the effect of any treatment at the TA level on students outcomes will depend not only on when the TAs were able to adjust their teaching behavior in response to feedback, but also on how important the TAs are for the students' performance (what is the weight of TAs in the education production function) - relative to other factors such as the students' effort or ability. The results on students' grades are, therefore, sensitive to several factors. Moreover, the teaching skills of the TA are only relevant for those students attending to TA's section, which is typically a smaller proportion of the total students enrolled in the section; however, since it is impossible to verify attendance of students without affecting their behavior, we are only able to use the information on grades for all the students enrolled in the section instead of using the grades only for those indeed attending it.[34]

If the intervention did in fact provide TAs with valuable and actionable feedback, we might expect to see a larger impact on outcomes in the following quarter (Winter 2013), when the treated TAs incorporated the Fall suggestions since the beginning of their teaching duties. We also evaluate the impact of the program during the Spring 2013 Quarter - to assess the potential persistence of the intervention.

Table 7 contains the regression analysis of the ITT and ToT effects pooling the Winter 2013 and Spring 2013 TAs' evaluations. Column 1 presents the mean difference between samples clustering errors at the TA level (most TAs teach two sections), corresponding to the

[33]The covariates included are age, male, English native speaker, PhD year, masters degree, number of quarters that they have taught, whether they have taught the course before and whether they have met with the TA coordinator.

[34]Another difficulty in the interpretation of the results on undergraduate grades is that undergraduate students might switch TA sessions within the same course, depending on the time of the discussion section and on the TA -for instance, if they believe that the TA does not meet their expectations. There is no way to account for this problem, the evaluations are anonymous and there is no way to track the switching. This measurement issue may reduce our ability to observe any effect of the treatment on the undergraduate students' performance. However, it should also be noted that the switching of sections tends to occur early in the quarter (if there is space available in another section) and, once they have chosen a section, the students attend that same section for the rest of the quarter.

results of table 5. In column 2 we add observable controls for the covariates. As expected, the magnitudes of the estimated coefficients did not change - we showed in section 3.1 that these variables are balanced between groups and should not affect our estimation, the standard errors were slightly reduced.[35]

Columns 3, 4 and 5 add controls for field of concentration, TAs' nationality and course fixed-effects, respectively. These covariates are not balanced between groups due to sample size constraints. Adding controls for the field of concentration of the TA does not change the results: the magnitude of the ITT effect is slightly larger, 0.37 points, and statistically significant for Winter 2013. The interaction term still shows that the effect is reduced for Spring to 0.235 points, although this interaction term is still not significantly different from zero. Including course fixed effects restricts the identification of the effect on those courses with various sections and it does change the magnitude. According to this result, the effect of peer feedback was almost three-fourths of a standard deviation for Winter 2013 and statistically significant at the 1% level. The increase in the effect of the intervention may be explained by the fact that most of the courses that have many sections are introductory courses: as the intervention focused on improving communication and pedagogy (rather than, say, improving content or knowledge), we might expect larger gains in courses that present basic ideas to a broad range of students, rather than in more advanced topics offered to economics majors. The interaction term (spring*ITT) suggests that for Spring the effect was reduced to 0.46 points or one-half of a standard deviation, but once more, these effects are imprecisely estimated.

Including fixed effects for nationality also changes the estimated coefficients, the effect is only identified for individuals belonging to a country with high representation in the sample - China (35%) or the US (20%), who make up a larger proportion of the effective sample in this specification. The estimated coefficient is reduced to 0.28 points for Winter and 0.19 points for Spring, with neither being statistically different from zero.

Column 6 adds the Fall 2012 TA mean evaluation reducing slightly the magnitude of the coefficient, but remaining statistically significant.[36] These results also show that there is persistence in the TAs' mean evaluations, those who obtained higher grades during the Fall 2012 Quarter also obtained higher grades in the following quarters.

The results for the ToT pooling the Winter 2013 and Spring 2013 samples are very similar the results of the previous table. The point estimate remains unchanged, fluctuating between 0.36 and 0.39 points for Winter 2013, and they are statistically different from zero

[35]The controls added are age, gender, being an English native speaker, PhD year, having a masters degree prior to starting the PhD, number of quarters taught at the university, having taught the same course in the past and having had a meeting with the TA coordinator.

[36]Note that this specification may control away some of the treatment effect - if the insignificant positive effect seen in the Fall 2012 represents some small improvement due to the intervention, by controlling for the Fall 2012 evaluation scores we restrict ourselves to examining the incremental improvement in outcomes between the Fall 2012 and Winter 2013 quarters, and Fall 2012 and Spring 2013 quarters, rather than the full impact of the intervention.

- focusing on columns 1, 2 and 6. The estimates for Spring 2013 are found between 0.171 and 0.187 points, although they are still imprecisely estimated.

Finally, table 7 also show results using the logarithm of the TAs' mean evaluations as the outcome measure. The results can be directly interpreted as percentage changes for the variables interest (ITT and ToT). Both the ITT and the ToT would suggest an increase between 4.5% and 4.9% in the average TAs' mean evaluations caused by the peer review program during Winter 2013 and between 2.56% and 3.36% during Spring 2013 (columns 1, 2 and 6 in both tables).[37]

3.3 Decomposition of TA evaluations

The evaluation forms used by the Office of Instructional Development have six areas that are assessed by the undergraduate students and that refer to concrete teaching skills: (i) the first category refers to the knowledge of the TA in the course taught; (ii) the second one evaluates the concern of the TA regarding the students understanding of the material; (iii) the third category focuses on the preparation and organization of the course; (iv) the fourth refers to the scope of the TA session relative to the course, more specifically whether the TA helped the students to improve their understanding on the material and expand on the topics covered in class; (v) the fifth area looks at the interaction between the TAs and the students outside the classroom; (vi) finally, the sixth component evaluates the communication skills of the TA, referring to the ability to transmit ideas. As before, all questions are evaluated in a scale from 1 (Very Low) to 9 (Very High).

Given that the overall TA evaluation is an assessment of all these categories, we believed that the intervention should have different impact across categories. In particular, we expected improvement in pedagogical areas which correspond to categories two, three, five and six, related to concern, organization, interaction and communication of the TA. It is unlikely that the intervention can change the knowledge of the TA or the scope of the sessions - most TAs follow instructions from professors about what topics to cover. However, if TAs are better organized and devote more time preparing their sections, undergraduate students might perceive an increase in the scope of the sections and/or in the knowledge of the TAs.

In this section we analyze the impact of the intervention in each of these areas using data from Winter 2013 and Spring 2013.[38] Table 8 shows the summary statistics for each of the six categories (knowledge, concern, organization, scope, interaction and communication).

[37] If we control for the course differences by adding course fixed effects, the ITT effect increases to 7% for Winter 2013 and 6.54% for Spring 2013, while the ToT remains relatively constant during Winter 2013 with a coefficient of 4.95% and a slight increase to 4.74% for Spring 2013. If we control by the differences in nationality (which may mask differences in teaching styles), the impact for Winter 2013 would be between 3.3% (ToT) and 3.9% (ITT) increase in TA evaluations, and between 1.4% (ToT) and 2.8% (ITT) for Spring 2013.

[38] We also analyzed the individual components of the TA evaluations for the Fall 2012 Quarter, but as with the overall TA evaluations, we found no significant effect of the intervention in any individual category (these results are available upon request).

Overall, students consider that the TAs of this department are knowledgable about the topics they are teaching (8.05 and 8.08 average grade for Winter 2013 and Spring 2013, respectively) and concerned about the students' learning (7.84 and 7.86 average grade for Winter 2013 and Spring 2013, respectively). However, the TAs seem to lag behind precisely in some of the areas at which the intervention is aiming at - organization, interaction and communication; the evaluations range from 7.37 to 7.76 average grades.

Table 9 presents the results of the regression analysis for each category, following the same specification as before, and again separately for the ITT and ToT groups. For the former, table 9 shows that with the exception of knowledge, all other components responded positively to the treatment. For Winter 2013, the estimated impact of the Intent to Treat ranges between 0.32 standard deviations (for organization) to 0.62 standard deviations (for communication). These results are statistically significant, with the exception of the estimated coefficient for organization. We also find that for concern, scope, organization, interaction and communication, the results are robust to the inclusion of covariates. For Spring 2013, we still find smaller magnitudes, ranging from 0.13 standard deviations (for organization) to 0.45 standard deviations (for communication). Interestingly, knowledge has a positive coefficient that would suggest an effect of 0.15 standard deviations, but not statistically different from zero. Similar to the results in section 3.2, the estimated coefficients are statistically significant for Winter 2013, while the interaction term that captures the effect for Spring 2013 is imprecisely estimated, but all point estimates are consistent with our previous results.

Overall, the intervention seems to have an important effect on the communication skills of the TAs, a smaller effect on organization, scope, concern and interaction. These result suggest that after the intervention TAs were more concerned about how they expressed themselves and how they could convey the material in a clearer manner.

4 Complementary Analyses

This section presents three complementary analyses. First, we study the effect on "average course evaluation." This measure is calculated using the grade that undergraduate students assign for the course in general (as opposed to the discussion section or the TA), but using only the information from students of a specific section (each section has capacity for 30 students). This figure might be considered a mixture of student satisfaction with the course, the professor and the TA.

The second part analyses attrition in a systematic way, analyzing who is offered (accepted) a TAship for Winter 2013 and Spring 2013. The TAships are allocated by the department based on the number of sections offered each quarter; however, graduate students might decline accepting the TAship if they receive an alternative funding source, for instance, a research assistant fellowship or a scholarship awarded by the Graduate Division

or an external organization.[39]

And the last part presents the main findings of the qualitative surveys that the treated TAs completed shortly after the intervention. The responses provide valuable information on which aspects of the intervention worked reasonably well and which aspects could be improved in the future.

4.1 Additional Analysis and Robustness

Table 4 shows the summary statistics for the average and median course evaluation by section for each quarter. For Fall 2012, the average course evaluation was is slightly lower than the average TA evaluation, 7.6 (in a 1 to 9 scale). These figures do not change for Winter 2013 and slightly improve for Spring 2013, for the latter it increased to 7.7.

Tables 10 presents the difference of means for the course average and course median for the three quarters and the two groups, ITT and ToT. As shown in Table 10, for Fall 2012, peer feedback had no effect on the average evaluation of the course, the magnitude is around 0.01 points and it is not statistically different from zero. For the median, the estimated coefficient is negative but small, -0.09 points, and still not statistically significant. In the case of the ToT, the magnitudes are slightly larger, remaining small and not statistically significant.

Once again, the results for Winter 2013 and Spring 2013 are larger in magnitude than those for the Fall 2012, but only statistically significant for Winter 2013. We estimate an effect of 0.47 points for the Winter 2013 mean evaluation, which represents almost one-half of a standard deviation. For the Spring 2013, the estimated effect is 0.21 points - or one-fifth of a standard deviation in that quarter.

Table 11 shows the results of the regression analysis adding controls. Column 1 shows the difference of means, column 2 to 6 add the controls used in section 3.2. In addition, column 3 controls for field fixed effects, column 4 for course fixed effects, column 5 for nationality fixed effects and, finally, column 6 adds the mean TAs' evaluation.[40] For Fall 2012 (panel A), the estimated coefficients become negative and larger in magnitude once we add controls; however, they remain not statistically different from zero - with the exception of the coefficients in column 6, which are statistically significant for the ITT estimates. These results are interesting, nonetheless, they suggest that despite the results on the mean TAs' evaluations (no effect was found), the presence of observers might have unintended negative effects among undergraduates. Interestingly, such potentially negative effects are

[39] Graduate students have incentives to look for an alternative source of funding. Acting as teaching assistants helps them to improve presentation and communication skills; however, becoming research assistants might help them to develop specific skills for research or they might be offered a co-authorship with a professor. Alternatively, seeking a scholarship from an independent organization might be done with the objective of freeing up time for research.

[40] The controls added are age, gender, being an English native speaker, PhD year, having a masters degree prior to starting the PhD, number of quarters taught at the university, having taught the same course in the past and having had a meeting with the TA coordinator.

not reflected on the evaluation of the TA, but on the evaluation of the course per se. A possible explanations is that the undergraduate students were aware of the presence of another TA and associate it with a negative signal for the quality of the course. A second interesting result is found in the last column, which shows that there is almost a one-to-one relationship between the average course evaluation and the average TA evaluation.

The analysis of pooling Winter 2013 and Spring 2013 are found in panel B of table 11. Each column adds controls as described in the previous paragraph. The ITT and ToT results for Winter 2013 range from one-half to three-quarters of a standard deviation and are significant at a 5% level, and the last column shows that the effect disappears once we add a control for the average TAs' evaluation, suggesting that the program boosts students' satisfaction through increasing TAs' performance. For Spring 2013, the estimated magnitudes range from one-third to three-sevenths of a standard deviation for the ITT coefficients, and between one-fifth and almost one-third of a standard deviation for the ToT coefficients. However, they remain imprecisely estimated - they are not statistically different from zero. Although this question (average course evaluation) is not intended to specifically measure students' satisfaction, it might reflect a weighted average of course quality, teaching ability of the main instructor and the performance of the TA. As the TA improves her skills, students might be able to reap the benefits of the course and associate this improved performance with higher overall quality of the class. These results are novel in this literature by avoiding potential problems of reverse causality - TA grades and overall evaluations are generally correlated in data sets, but it might be because better instructors increase the perceived performance of the TAs; our experiment does not affect anything else in the class except the TAs.

We turn now to analyze one alternative measure of performance - the probability of meeting with the TA coordinator. The results from section 3 show that the peer review program was successful one quarter after the program took place (Winter 2013) and suggest some persistence of effects two quarters later (Spring 2013), although these effects decline in magnitude and are imprecisely estimated. The positive impact of the program on average TAs' evaluations suggest that it might have affected the likelihood of meeting with the TA coordinator. The Department requires that TAs obtaining an average grade below seven - in at least one of their sections - meet with the TA coordinator to discuss strategies for improving their teaching practices.

Finally, table 12 analyzes how the probability of obtaining an average grade below seven (in at least one section) is affected by the program. We consider this measure an alternative indicator of the success of the peer review program. The coefficients are very imprecisely estimated, but they are consistent with the results of the previous section. The ITT results suggest that the program decreased the probability of having an average grade below seven by 5.9 percentage points for Winter 2013 and by 1.9 percentage points for Spring 2013. The

ToT results are similar, the program decreased this probability by 6.2 and 1.6 percentage points for Winter 2013 and Spring 2013, respectively.

4.2 Attrition

Attrition during Winter 2013 and Spring 2013 might be a concern if it affects differentially the treatment and the control groups. However, it might also be an outcome of the program itself. Treated TAs might exert an effort to continue teaching if they believe that they can reap the benefits of the feedback received. Alternatively, they might been offered the slot because of better performance or the Department might have favored these students by offering them the TAships first - believing that the program was useful for them.[41]

Table 13 shows the attrition rate by quarter for the two quarter after Fall 2012. From our original sample (55 teaching assistants randomly assigned to treatment and control groups), we lost 7.3% of them in Winter 2013 and 12.7% in Spring 2013.

In table 14 we present the results of a linear probability model estimating if the attrition rate differed by treatment status.[42] The results support the hypothesis that treated TAs (or assigned to treatment) are indeed less likely to stop teaching in the next two quarters. The coefficients are not statistically significant but they suggest that these TAs were 10 percentage points more likely to still be teaching in Winter 2013 and 0.5 percentage points of still be teaching in Spring 2013.

4.3 Feedback on the intervention from participants

Finally, we conducted a qualitative survey at the end of the Fall 2012 Quarter in order to assess some of the key aspects of the intervention. In particular, we assessed two areas: (i) The first set of questions tried to elicit information regarding the experience of the TAs while they were being observed; (ii) the second set of questions tried to gather information on the TAs in their role as observers and evaluators.

For the first part, our objective was to qualitatively assess whether the TAs were aware (self-conscious) of the presence of observers, whether they modified their behavior while teaching or preparing for class, and whether they found the feedback received useful. Overall, the project ran smoothly, all TAs claim to have received their feedback shortly after being observed and all TAs claim to have read it carefully. Most of the TAs agreed that the feedback was useful; however, they believe that it mostly contained positive reinforcement or that it pointed out problems of which the TAs were already aware, instead of pointing specific actions they could take in order to improve their teaching. Regarding the effect of

[41] TAships for the following quarter are usually distributed during the last two weeks of a given quarter, therefore, they were allocated before the research team had access to the TA evaluations (they are processed by the Office of Instructional Development and distributed back to the Departments during the third/fourth week of the following quarter) for the assessment of the peer review program and, thus, before the Department knew whether it was successful or not.

[42] Similar results are found using a probit model.

the observation itself, the responses were mixed: some TAs did not notice the observers, while others felt somewhat uncomfortable while they were being observed.

Despite self-awareness, there are mixed responses regarding their attitude towards the preparation of the class: a few TAs modified their behavior knowing that a fellow TA could be present. Finally, we also included a question regarding what type of observer the TAs would rather have (peers or experts) - our concern was that since all observers were other graduate students of the same department, the TAs would feel more nervous or uncomfortable compared to having a stranger observing them. Even though the responses to this question are also mixed, a majority prefers having somebody within the department evaluating their classes.

The objective of the second part of the survey was designed to evaluate the perception of the TAs regarding their qualifications as evaluators. The first question, which referred to the initial training workshop, showed that most of the TAs did not believe that it helped them to improve their abilities as evaluators. This is an important point to consider for future interventions: the training workshop should be carefully tailored to the needs of each department and the contents should be revised or it might be eliminated altogether. The next question tried to assess whether they felt capable of performing the task. Most of them agreed that they could evaluate the teaching skills of their colleagues and that the feedback format helped them to transmit their thoughts and comments. Regarding the perceived change in teaching "skills" between the two observations, most TAs did not feel that there was any improvement, which is consistent with the results for the Fall 2012 Quarter. We were also interested in knowing if they would feel more comfortable evaluating a stranger from a different department, but most of the answers expressed a preference for observing TAs in the same Department.

The last question elicited their perception of the project overall. In particular, we were interested in knowing whether they believed that it was useful and had potential for a large scale implementation. 80% of them answered that they liked the project and that they believed that it had potential. Overall, the TAs took their role seriously during this experiment, they considered themselves fit to observe and assess the teaching skills of their peers, and they preferred both to observe and to be observed by peers of the same department.

5 Conclusions

Peer feedback is a potentially attractive intervention for improving teaching practices; even though some studies of the educational literature have assessed its effectiveness (mostly from a qualitative perspective), to the best of our knowledge, there is no study with a quantitative approach that allows to evaluate the program in terms of causality. Our study is a first step to fill this gap in the literature by using a randomized intervention in one Department of a large public university to establish a causal relationship between peer

feedback and teaching performance of TAs.

The results from the study suggest that peer feedback at the TA level had a positive but not significant effect on the overall TA student evaluations during the quarter that the intervention took place (Fall 2012). The RCT, however, had a large effect in the following two quarters: it increased the TAs evaluations by one half of a standard deviation during Winter 2013 and by one-fourth of a standard deviation in Spring 2013.

In terms of the specific areas of improvement, the results show that the intervention improved communication skills, and that it had a smaller, less significant effect on organization, scope, concern and interaction with the students. As expected, the intervention had no effect on how knowledgeable about the material covered in the section was the TA.

Finally, we show a novel result using the average course evaluation as a proxy of students' satisfaction with th-e course. The average course evaluation is the mean evaluation for the course calculated using only the information of the students in a given section. The results show that the peer review program increased the average course evaluation by one-half of a standard deviation for Winter 2013 and by one-fifth of a standard deviation for Spring 2013. These results suggest that the performance of TAs has a causal positive impact on the satisfaction of undergraduate students with their courses.

Finally, we conducted a qualitative survey which provided valuable information on the components of the intervention that worked better and those which must be improved for future interventions. Regarding the first aspect, the TAs felt comfortable having other graduate students from the same department observing them (as opposed to having a stranger as evaluator), and they indicate that the observation and feedback formats help them to better assess their peers. On the other hand, they did not find that training workshop held at the beginning of the quarter was helpful (on teaching them how to provide feedback) and the feedback received did not contain enough specific actions that they could take for improving their teaching. However, overall, they found the program valuable.

6 Main Figures and Tables

Figure 1: Observation Format

TA being observed: _____

Time and place:_____

Observer:_____

OBSERVATION FORMAT

INSTRUCTIONS:

Read the format before attending the session, so that you know what to look for. Make sure to know what topics and concepts are going to be covered during the session beforehand.

		Not aplicable	Strongly disagree		Disagree		Neither agree nor disagree		Agree		Strongly agree
	Organization/Clarity										
1	The aims, objectives and structure of the session were clear.	N/A	1	2	3	4	5	6	7	8	9
2	The topic and concepts covered were prepared beforehand.	N/A	1	2	3	4	5	6	7	8	9
3	The TAs speech was easy to understand.	N/A	1	2	3	4	5	6	7	8	9
4	The board or other teaching aids were used appropriately.	N/A	1	2	3	4	5	6	7	8	9
5	The TA managed properly the time of the session	N/A	1	2	3	4	5	6	7	8	9
	Specific comments on this factor:										
	Group Interaction										
6	The TA effectively managed the group interaction.	N/A	1	2	3	4	5	6	7	8	9
7	The TA encouraged students to actively participate in the session.	N/A	1	2	3	4	5	6	7	8	9
8	Students were engaged in the explanation and discussion of the section.	N/A	1	2	3	4	5	6	7	8	9
	Specific comments on this factor:										
	Instructor Enthusiasm										
9	The TA was enthusiastic about and interested in the topic.	N/A	1	2	3	4	5	6	7	8	9
10	The TA developed good rapport with the students and responded to their needs.	N/A	1	2	3	4	5	6	7	8	9
	Specific comments on this factor:										

	Learning/Value										
11	The TA explained things well and the examples used helped the students to understand the topic.	N/A	1	2	3	4	5	6	7	8	9
12	Ideas were transmitted clearly and in a way students would understand them.	N/A	1	2	3	4	5	6	7	8	9
13	The TA's feedback/answers to questions helped students to learn.	N/A	1	2	3	4	5	6	7	8	9
	Specific comments on this factor:										
	Breadth of Coverage										
14	The session was well integrated with the rest of the course (following the syllabus).	N/A	1	2	3	4	5	6	7	8	9
15	The concepts discussed were framed into the broad scope of the course.	N/A	1	2	3	4	5	6	7	8	9
16	The TA linked the topics in a coherent manner.	N/A	1	2	3	4	5	6	7	8	9
	Specific comments on this factor:										
	Comments										
17	Please list the three best things about the TA.										
18	Please list three suggestions for improving the session.										
19	Comments on the lesson plan e.g. activities, structure and timing.										

Note: Based on the Danielson framework (Danielson, 2011) of assessing teaching skills and also, based on the students' evaluations used in the large public university. It is also consistent with the framework developed by Marsh (1983) on what set of factors are important for good teaching.

Figure 2: Feedback Format

TA being observed: _____

Time and place:_____

Observer:_____

FEEDBACK FORMAT

INSTRUCTIONS:

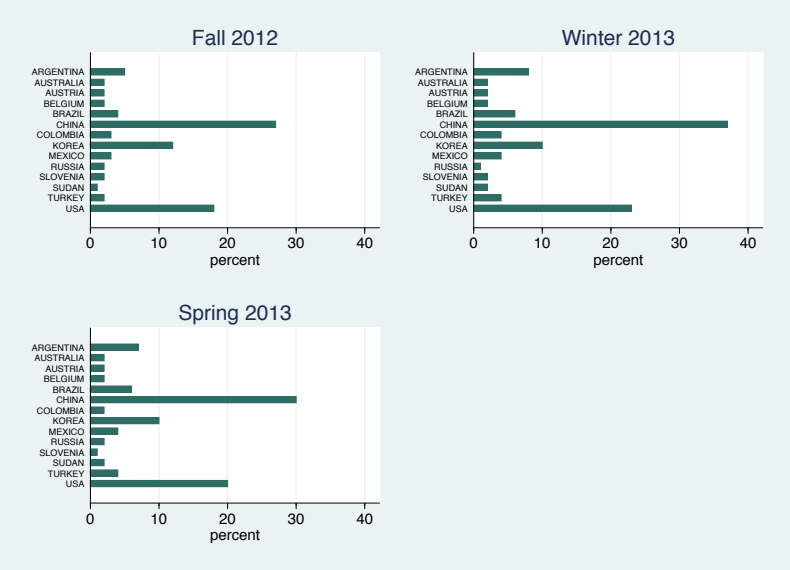
Please record the main comments and feedback points you would like to provide to your peer TA. Please be very specific about the actions she can take in each field to improve her performance.

1	*Organization/Clarity*
	Specific comments on this factor:
	Specific actions towards improvement:
2	*Group Interaction*
	Specific comments on this factor:
	Specific actions towards improvement:
3	*Instructor Enthusiasm*
	Specific comments on this factor:
	Specific actions towards improvement:
4	*Learning/Value*
	Specific comments on this factor:
	Specific actions towards improvement:

5	*Breadth of Coverage*
	Specific comments on this factor:
	Specific actions towards improvement:
6	Comments
	Other comments

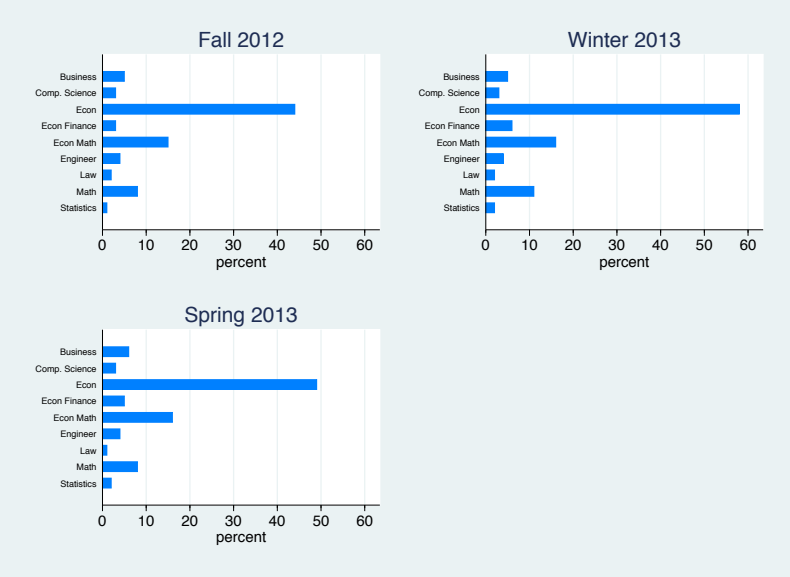
Note: Based on the Danielson framework (Danielson, 2011) of assessing teaching skills and also, based on the students' evaluations used in the large public university. It is also consistent with the framework developed by Marsh (1983) on what set of factors are important for good teaching.

Figure 3: TAs Nationality



Note: This figure shows the country of origin of all TAs as reported by them in the baseline survey.

Figure 4: TAs Undergraduate Major



Note: This figure shows the TAs undergraduate major as reported by them in the baseline survey.

Figure 5: TAs PhD Main Field

Note: This figure shows the TAs main Field of specialization in the PhD as reported by them in the baseline survey.

Table 1: Descriptive Statistics: Covariates

Panel A. Fall					
Variable	**Mean**	**Median**	**St. Dev.**	**Min**	**Max**
Age	27.02	27	2.32	23	32
I(male)	0.75	1	0.44	0	1
I(English native)	0.22	0	0.42	0	1
PhD year	3.15	3	0.89	2	5
I(MA)	0.40	0	0.49	0	1
Quarters taught	6.40	6	4.65	0	18
I(taught this course before)	0.45	0	0.50	0	1
I(coordinator)	0.14	0	0.37	0	1
Panel B. Winter					
Variable	**Mean**	**Median**	**St. Dev.**	**Min**	**Max**
Age	27.00	27	2.31	23	32
I(male)	0.76	1	0.43	0	1
I(English native)	0.22	0	0.42	0	1
PhD year	3.06	3	0.83	2	5
I(MA)	0.41	0	0.50	0	1
Quarters taught	7.12	7	4.34	1	18
I(taught this course before)	0.57	1	0.50	0	1
I(coordinator)	0.34	0	0.39	0	1
Panel C. Spring					
Variable	**Mean**	**Median**	**St. Dev.**	**Min**	**Max**
Age	27.10	27	2.28	23	32
I(male)	0.75	1	0.44	0	1
I(English native)	0.23	0	0.42	0	1
PhD year	3.10	3	0.88	2	5
I(MA)	0.44	0	0.50	0	1
Quarters taught	7.31	7	4.66	1	19
I(taught this course before)	0.63	1	0.49	0	1
I(coordinator)	0.33	0	0.39	0	1

Note: This table presents the descriptive statistics of the observable characteristics of the TAs as reported by them in the baseline survey.

Table 2: Randomization by Course (Number of TAs)

	Course	Control	ITT	Treatment
Principles of Economics	Econ 1	2	3	2
Principles of Economics	Econ 2	3	3	2
Microeconomic Theory	Econ 11	5	7	5
Statistics for Economists	Econ 41	2	5	5
Microeconomic Theory	Econ 101	2	3	3
Macroeconomic Theory	Econ 102	2	2	2
Introduction to Econometrics	Econ 103L	2	3	3
Economics of Technology and E-commerce	Econ 106TL	1	2	0
Investments	Econ 106VL	1	2	1
Public Economics	Econ 130L	0	1	1
Economic Growth	Econ 164L	0	1	1
Microeconomic Theory (Grad)	Econ 201A	1	0	0
Macroeconomic Theory (Grad)	Econ 202A	1	0	0
Econometrics (Graduate)	Econ203A	1	0	0
	Total	23	32	25

Note: The table displays the courses offered by the Department that have TAs. Most of the introductory courses offered have various TAs, and we stratified the randomization accordingly when possible.

Table 3: Balancing of Covariates

Panel A. Fall						
Variable	**Control**	**ITT**	**p-value (equal means)**	**Control**	**ToT**	**p-value (equal means)**
Age	27.04	27.00	[0.946]	27.04	26.84	[0.764]
I(male)	0.70	0.78	[0.481]	0.70	0.76	[0.625]
I(english native)	0.13	0.28	[0.188]	0.13	0.28	[0.211]
PhD year	3.13	3.16	[0.917]	3.13	3.04	[0.716]
I(MA)	0.52	0.31	[0.123]	0.52	0.32	[0.163]
Quarters taught	6.39	6.41	[0.991]	6.39	6.04	[0.784]
I(taught this course before)	0.43	0.47	[0.807]	0.43	0.48	[0.760]
I(coordinator)	0.17	0.16	[0.864]	0.17	0.16	[0.900]
N	23	32		23	25	
Panel B. Winter						
Variable	**Control**	**ITT**	**p-value (equal means)**	**Control**	**ToT**	**p-value (equal means)**
Age	27.00	27.00	[1.000]	27.00	26.84	[0.817]
I(male)	0.75	0.77	[0.846]	0.75	0.76	[0.940]
I(english native)	0.15	0.26	[0.370]	0.15	0.28	[0.308]
PhD year	2.95	3.13	[0.460]	2.95	3.04	[0.701]
I(MA)	0.55	0.32	[0.111]	0.55	0.32	[0.126]
Quarters taught	6.40	7.58	[0.348]	6.40	7.04	[0.594]
I(taught this course before)	0.50	0.61	[0.437]	0.50	0.56	[0.697]
I(coordinator)	0.20	0.16	[0.730]	0.20	0.16	[0.734]
N	20	31		20	25	
Panel C. Spring						
Variable	**Control**	**ITT**	**p-value (equal means)**	**Control**	**ToT**	**p-value (equal means)**
Age	27.30	26.96	[0.620]	27.30	26.83	[0.491]
I(male)	0.70	0.79	[0.509]	0.70	0.74	[0.782]
I(english native)	0.15	0.29	[0.280]	0.15	0.30	[0.242]
PhD year	3.05	3.14	[0.723]	3.05	3.09	[0.886]
I(MA)	0.60	0.32	[0.057]	0.60	0.30	[0.053]
Quarters taught	6.95	7.57	[0.654]	6.95	7.22	[0.844]
I(taught this course before)	0.55	0.68	[0.375]	0.55	0.61	[0.705]
I(coordinator)	0.20	0.18	[0.855]	0.20	0.17	[0.831]
N	20	28		20	23	

Note: This tables depicts the summary statistics of the observable characteristics of the TA participating in the intervention: age, indicator variable for male, indicator variable for being an English Native, the PhD year the TA is currently attending to, an indicator variable for obtaining a Masters Degree before entering the PhD, number of quarters as a TA in the current university, an indicator variable of whether the TA has taught the course before, and finally and indicator variable of whether the TA has been called by the TA coordinator of the Department due to obtaining very low scores in previous students' evaluations.

Table 4: Descriptive Statistics: Outcome Variables

Panel A. Fall						
Variable	**Mean**	**Median**	**St. Dev.**	**Min**	**Max**	**N**
Average evaluation of TA	7.86	7.95	0.74	5.67	9	99
Median evaluation of TA	8.20	8.00	0.73	6.00	9	99
Average evaluation of course	7.64	7.77	0.81	5.00	9	96
Median evaluation of course	8.09	8.00	0.90	5.50	9	96
Panel B. Winter						
Variable	**Mean**	**Median**	**St. Dev.**	**Min**	**Max**	**Max**
Average evaluation of TA	7.81	7.99	0.79	5.10	8.9	94
Median evaluation of TA	8.09	8.00	0.95	5.00	9	94
Average evaluation of course	7.65	7.77	0.88	4.60	9	94
Median evaluation of course	7.95	8.00	1.04	5.00	9	94
Panel C. Spring						
Variable	**Mean**	**Median**	**St. Dev.**	**Min**	**Max**	**Max**
Average evaluation of TA	7.88	8.17	0.96	4.25	8.88	85
Median evaluation of TA	8.10	8.00	1.03	5.00	9	85
Average evaluation of course	7.72	8.05	1.13	2.00	8.92	85
Median evaluation of course	7.99	8.00	1.34	1.50	9	85

Note: This table presents the descriptive statistics of the average and median TAs' and course evaluations, both calculated using only the undergraduate students of a specific TA section.

Table 5: Difference of Means: ITT and ToT

Panel A. Intent to Treat **Fall** **Variable**	**Control**	**ITT**	**Difference**	**p-value** **(equal means)**
Average evaluation of TA	7.76	7.92	0.17	[0.277]
Median evaluation of TA	8.15	8.22	0.08	[0.610]
Grade (dev. from course mean)	0.004	0.003	-0.001	[0.769]
N	41	58		
Winter				
Average evaluation of TA	7.58	7.95	0.37	[0.029]**
Median evaluation of TA	7.84	8.24	0.39	[0.051]*
N	35	59		
Spring				
Average evaluation of TA	7.74	7.98	0.24	[0.253]
Median evaluation of TA	7.93	8.22	0.29	[0.209]
N	34	51		
Panel B. Treatment on the Treated **Fall** **Variable**	**Control**	**ToT**	**Difference**	**p-value** **(equal means)**
Average evaluation of TA	7.76	7.92	0.16	[0.334]
Median evaluation of TA	8.15	8.20	0.06	[0.723]
Grade (dev. from course mean)	0.004	0.001	-0.003	[0.565]
N	41	44		
Winter				
Average evaluation of TA	7.58	7.94	0.36	[0.043]**
Median evaluation of TA	7.84	8.21	0.37	[0.096]*
N	35	48		
Spring				
Average evaluation of TA	7.74	7.92	0.19	[0.406]
Median evaluation of TA	7.93	8.16	0.23	[0.340]
N	34	41		

Note: This table shows the difference of the averages of the main outcome variables of interest between the control group and the ITT group (TAs who were offered to participate in the program) and the treated group (TAs who actually participated in the program). The last column shows the p-value of the test of equality of mean between two groups.. The outcomes of interest are the TAs overall evaluation and the deviation of the section average grade from the course average (recall that most of the courses had many sections). The last column shows the p-value of the test of equality of mean between two groups.

Table 6: Regression Analysis: TA Evaluation

	Independent Variable: Fall Average Course Evaluations				
	(1)	(2)	(3)	(4)	(5)
Average Course Evaluations					
Intent to Treat	0.1441	0.0458	0.1232	-0.0679	-0.1769
	[0.182]	[0.182]	[0.169]	[0.172]	[0.221]
ToT	0.1303	-0.0129	-0.0126	-0.2364	-0.2151
	[0.196]	[0.205]	[0.203]	[0.186]	[0.226]
Grade (dev. from course mean)					
Intent to Treat	-0.0018	-0.0064	-0.0045	-0.0054	-0.0072
	[0.005]	[0.005]	[0.005]	[0.005]	[0.006]
ToT	-0.0037	-0.0077	-0.0068	-0.0077	-0.0392
	[0.006]	[0.005]	[0.006]	[0.007]	[0.024]
Log (Average Course Evaluations)					
Intent to Treat	0.0198	0.0056	0.0161	-0.0092	-0.0258
	[0.025]	[0.024]	[0.022]	[0.023]	[0.030]
ToT	0.0180	-0.0022	-0.0010	-0.0312	-0.0310
	[0.026]	[0.027]	[0.027]	[0.025]	[0.030]
Control variables	No	Yes	Yes	Yes	Yes
Field of concentration dummies	No	No	Yes	No	No
Course dummies	No	No	No	Yes	No
Nationality dummies	No	No	No	No	Yes

Robust standard errors in brackets and clustered by TA.
*** p<0.01, ** p<0.05, * p<0.1

Note: Results from OLS regressions in which the dependent variable is the TA's average student overall evaluation by section (in general, TAs are responsible for teaching two sections) and the variables of interest are the ITT, an indicator variable of the intent to treat, and the ToT, an indicator variable of the treatment on the treated. Columns 2 to 5 control for age, male, English native speaker, masters degree before the PhD, number of quarters taught, a variable indicating if the TA has taught the same course before, and an indicator variable for having met the TA coordinator. Column 3 controls for field of concentration fixed effects, column 4 for course fixed effects and column 5 for nationality fixed effects. Robust errors cluster by TA.

Table 7: Regression Analysis Winter and Spring 2013: TA Evaluation

	Independent Variable: Winter and Spring Average Evaluations of TAs					
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Intent to Treat						
Intent to Treat	0.3661*	0.3624**	0.3765**	0.5545***	0.3488	0.3599**
	[0.194]	[0.179]	[0.182]	[0.187]	[0.209]	[0.163]
Spring	0.1556	0.1088	0.0996	0.1401	0.0948	0.1059
	[0.192]	[0.212]	[0.214]	[0.265]	[0.228]	[0.220]
ITT*Spring Quarter	-0.1222	-0.0889	-0.0958	-0.1073	-0.0770	-0.1034
	[0.218]	[0.235]	[0.233]	[0.296]	[0.249]	[0.242]
Panel B. ToT						
ToT	0.3631*	0.3774**	0.3540*	0.4398**	0.3076	0.3857**
	[0.207]	[0.187]	[0.204]	[0.202]	[0.217]	[0.172]
Spring	0.1556	0.1075	0.0963	0.0762	0.1042	0.1016
	[0.192]	[0.214]	[0.218]	[0.288]	[0.231]	[0.223]
ToT*Spring Quarter	-0.1757	-0.1510	-0.1535	-0.1430	-0.1567	-0.1624
	[0.226]	[0.240]	[0.243]	[0.328]	[0.255]	[0.249]
Panel C. Intent to Treat: Log Evaluations of TAs						
Intent to Treat	0.0490*	0.0493**	0.0515**	0.0729***	0.0488*	0.0495**
	[0.027]	[0.024]	[0.025]	[0.026]	[0.028]	[0.023]
Spring	0.0167	0.0095	0.0082	0.0118	0.0075	0.0092
	[0.028]	[0.032]	[0.032]	[0.040]	[0.034]	[0.033]
ITT*Spring Quarter	-0.0154	-0.0103	-0.0111	-0.0096	-0.0085	-0.0119
	[0.031]	[0.034]	[0.034]	[0.044]	[0.036]	[0.035]
Panel D. ToT: Log Evaluations of TAs						
ToT	0.0482*	0.0514**	0.0495*	0.0561*	0.0426	0.0530**
	[0.029]	[0.025]	[0.028]	[0.028]	[0.029]	[0.024]
Spring	0.0167	0.0093	0.0077	0.0027	0.0087	0.0086
	[0.028]	[0.032]	[0.033]	[0.044]	[0.035]	[0.033]
ToT*Spring Quarter	-0.0219	-0.0184	-0.0185	-0.0128	-0.0191	-0.0197
	[0.032]	[0.035]	[0.035]	[0.048]	[0.037]	[0.036]
Control variables	No	Yes	Yes	Yes	Yes	Yes
Field of concentration dummies	No	No	Yes	No	No	No
Course dummies	No	No	No	Yes	No	No
Nationality dummies	No	No	No	No	Yes	No
Fall Average TA evaluation	No	No	No	No	No	Yes

Robust standard errors in brackets and clustered by TA.
*** p<0.01, ** p<0.05, * p<0.1

Note: Results from OLS regressions in which the dependent variable is the average TA evaluation per section. The variables of interest are ITT and ToT, indicator variables of the intent to treat and of the treatment on the treated, respectively. And the interaction terms of the spring quarter with the ITT and the ToT, respectively. Column 1 shows the simple means, column 2 to 6 add the controls set to pre-treatment levels - age, male, English native speaker, masters degree before the PhD, number of quarters taught, a variable indicating if the TA has taught the same course before, and an indicator variable for having met the TA coordinator. Column 3 controls for field fixed effects, column 4 for course fixed effects, column 5 for nationality fixed effects and, finally, column 6 adds the mean TA evaluation. Robust errors cluster by TA.

Table 8: Summary Statistics Winter and Spring 2013: Other outcomes

Average evaluation of TA's	Mean	Median	Std. Dev.	Min	Max
Panel A. Winter					
Knowledge	8.05	8.22	0.62	5.83	8.95
Concern	7.84	7.91	0.74	5.50	9.00
Organization	7.75	7.90	0.78	4.80	8.90
Scope	7.65	7.77	0.76	5.20	8.88
Interaction	7.76	7.84	0.76	5.67	8.95
Communication	7.37	7.50	1.04	4.00	8.89
Panel B. Spring					
Knowledge	8.08	8.25	0.70	5.78	8.96
Concern	7.86	8.10	0.90	4.00	8.93
Organization	7.81	8.11	0.97	4.00	8.92
Scope	7.70	7.84	0.99	3.00	8.91
Interaction	7.83	8.14	0.93	3.50	8.93
Communication	7.51	7.92	1.20	3.00	8.91

Note: The table presents the summary statistics of the more specific questions of the students' evaluation of the TA regarding how knowledgeable the TA is, how concern is the TA about the students' learning, the organization and preparation of the section, the scope of the section, how welcome students felt (interaction), and the TAs' communication skills.

Table 9: Regression Analysis Winter and Spring 2013: Other outcomes

	Independent Variable: Winter and Spring Average Evaluations of TAs					
	Knowledge		Concern		Organization	
	(1)	(2)	(3)	(4)	(5)	(6)
Intent to Treat	**0.0739**	**0.0069**	**0.2898***	**0.2622**	**0.3131**	**0.2516**
	[0.141]	[0.151]	[0.165]	[0.170]	[0.187]	[0.193]
Spring	-0.0255	-0.0319	0.0992	0.0992	0.1446	0.1313
	[0.148]	[0.151]	[0.183]	[0.181]	[0.159]	[0.164]
ITT*Spring	**0.0862**	**0.0993**	**-0.1283**	**-0.1135**	**-0.1414**	**-0.1214**
	[0.181]	[0.186]	[0.210]	[0.208]	[0.193]	[0.195]
Treatment in the treated						
ToT	**0.0596**	**-0.0055**	**0.2758**	**0.2476**	**0.3444**	**0.2829**
	[0.158]	[0.161]	[0.181]	[0.175]	[0.206]	[0.204]
Spring	-0.0255	-0.0302	0.0992	0.1021	0.1446	0.1324
	[0.148]	[0.149]	[0.183]	[0.178]	[0.160]	[0.163]
ToT*Spring	**0.0635**	**0.0769**	**-0.1945**	**-0.1840**	**-0.2063**	**-0.1921**
	[0.194]	[0.196]	[0.217]	[0.209]	[0.204]	[0.201]
	Scope		**Interaction**		**Communication**	
	(7)	**(8)**	**(9)**	**(10)**	**(11)**	**(12)**
Intent to Treat	**0.3268***	**0.3098***	**0.3128***	**0.3007**	**0.6760****	**0.6493****
	[0.172]	[0.167]	[0.180]	[0.184]	[0.281]	[0.268]
Spring	0.0819	0.0833	0.1140	0.1135	0.2246	0.2252
	[0.195]	[0.197]	[0.187]	[0.188]	[0.228]	[0.220]
ITT*Spring	**-0.0392**	**-0.0318**	**-0.0686**	**-0.0571**	**-0.1170**	**-0.1064**
Treatment in the treated						
ToT	**0.3168***	**0.3071***	**0.3127**	**0.2959**	**0.6403****	**0.6029****
	[0.186]	[0.175]	[0.194]	[0.192]	[0.297]	[0.275]
Spring	0.0819	0.0843	0.1140	0.1156	0.2246	0.2293
	[0.195]	[0.194]	[0.187]	[0.185]	[0.228]	[0.215]
ToT*Spring	**-0.1110**	**-0.0971**	**-0.1283**	**-0.1141**	**-0.1431**	**-0.1189**
	[0.235]	[0.234]	[0.220]	[0.218]	[0.274]	[0.264]
Control Variables	No	Yes	No	Yes	No	Yes

Robust standard errors in brackets and clustered by TA
*** p<0.01, ** p<0.05, * p<0.1

Note: The table presents the results from the OLS regressions on each individual category evaluated by the undergraduate students: knowledge, concern, organization, scope, interaction and communication. The variables of interest are ITT and ToT, indicator variables of the intent to treat and of the treatment on the treated, respectively. And the interaction terms of the spring quarter with the ITT and the ToT, respectively. Columns 1, 3, 5, 7, 9 and 11 present the difference in means. Columns 2, 4, 6, 8, 10 and 12 control for age, male, English native speaker, masters degree before the PhD, number of quarters taught, a variable indicating if the TA has taught the same course before, and an indicator variable for having met the TA coordinator. Robust errors cluster by TA.

Table 10: Difference of Means of Average Course Evaluations: ITT and ToT

Panel A. Intent to Treat **Fall** **Variable**	**Control**	**ITT**	**Difference**	**p-value (equal means)**
Average evaluation of course	7.63	7.64	0.01	[0.962]
Median evaluation of course	8.15	8.05	-0.09	[0.617]
N	41	58		
Winter				
Average evaluation of course	7.35	7.82	0.47	[0.011]**
Median evaluation of course	7.61	8.15	0.54	[0.014]**
N	35	59		
Spring				
Average evaluation of course	7.54	7.84	0.31	[0.222]
Median evaluation of course	7.81	8.12	0.31	[0.299]
N	34	51		
Panel B. Treatment on the Treated **Fall** **Variable**	**Control**	**ToT**	**Difference**	**p-value (equal means)**
Average evaluation of course	7.63	7.69	0.06	[0.745]
Median evaluation of course	8.15	8.05	-0.10	[0.622]
N	41	44		
Winter				
Average evaluation of course	7.35	7.83	0.47	[0.018]**
Median evaluation of course	7.61	8.10	0.49	[0.040]**
N	35	48		
Spring				
Average evaluation of course	7.54	7.78	0.25	[0.358]
Median evaluation of course	7.81	8.05	0.24	[0.445]
N	34	41		

Note: This table shows the difference of the means of the average (median) course evaluation between the control group and the ITT group (TAs who were offered to participate in the program), and between the control group and the ToT group (TAs who actually participated in the program). The last column shows the p-value of the test of equality of mean between two groups.

Table 11: Regression Analysis: Average Course Evaluations

	Independent Variable: Average Course Evaluations					
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Fall						
Intent to Treat	0.0087	-0.1313	-0.0430	-0.2250	-0.3438	-0.1913**
	[0.201]	[0.192]	[0.168]	[0.205]	[0.229]	[0.072]
TA evaluation						0.9836***
						[0.051]
ToT	0.0610	-0.0808	-0.0217	-0.2299	-0.3270	-0.0885
	[0.220]	[0.227]	[0.238]	[0.233]	[0.243]	[0.068]
TA evaluation						0.9982***
						[0.048]
Panel B. Winter and Spring						
Intent to Treat	0.4706**	0.4593**	0.4749**	0.6196***	0.5077**	0.0624
	[0.211]	[0.199]	[0.201]	[0.220]	[0.223]	[0.062]
Spring	0.1816	0.1201	0.1136	0.1259	0.1078	0.0009
	[0.246]	[0.280]	[0.284]	[0.350]	[0.303]	[0.087]
ITT*Spring Quarter	-0.1638	-0.1200	-0.1299	-0.1279	-0.1203	-0.0226
	[0.279]	[0.310]	[0.310]	[0.384]	[0.328]	[0.101]
TA evaluation						1.0952***
						[0.048]
ToT	0.4712**	0.4818**	0.4671**	0.4983**	0.4450*	0.0652
	[0.228]	[0.210]	[0.231]	[0.245]	[0.235]	[0.067]
Spring	0.1816	0.1139	0.1060	0.0646	0.1147	-0.0048
	[0.246]	[0.284]	[0.290]	[0.384]	[0.306]	[0.088]
ToT*Spring Quarter	-0.2231	-0.1935	-0.1964	-0.1956	-0.2090	-0.0267
	[0.292]	[0.319]	[0.323]	[0.429]	[0.337]	[0.109]
TA evaluation						1.1037***
						[0.052]
Control variables	No	Yes	Yes	Yes	Yes	Yes
Field of concentration dummies	No	No	Yes	No	No	No
Course dummies	No	No	No	Yes	No	No
Nationality dummies	No	No	No	No	Yes	No
Fall Average TA evaluation	No	No	No	No	No	Yes

Robust standard errors in brackets and clustered by TA.
*** p<0.01, ** p<0.05, * p<0.1

Note: Results from OLS regressions in which the dependent variable is the average course evaluation per section - i.e. calculated using the subsample of undergraduate students belonging to a specific section. The variables of interest are ITT and ToT, indicator variables of the intent to treat and of the treatment on the treated, respectively. And the interaction terms of the spring quarter with the ITT and the ToT, respectively. Column 1 shows the simple means, column 2 to 6 add the controls - age, male, English native speaker, masters degree before the PhD, number of quarters taught, a variable indicating if the TA has taught the same course before, and an indicator variable for having met the TA coordinator. Column 3 controls for field fixed effects, column 4 for course fixed effects, column 5 for nationality fixed effects and, finally, column 6 adds the mean TA evaluation. Robust errors cluster by TA.

Table 12: Regression Analysis ITT and ToT: Grade below seven

	Independent Variable: Grade below seven					
	(1)	(2)	(3)	(4)	(5)	(6)
Intent to Treat						
ITT	**-0.0591**	**-0.0540**	**-0.0674**	**-0.1362**	**-0.0296**	**-0.0530**
	[0.119]	[0.111]	[0.113]	[0.145]	[0.126]	[0.104]
Spring Quarter	0.0067	0.0294	0.0338	0.0029	0.0346	0.0307
	[0.140]	[0.140]	[0.142]	[0.172]	[0.148]	[0.142]
ITT*Spring Quarter	**0.0395**	**0.0262**	**0.0299**	**0.0575**	**0.0129**	**0.0324**
	[0.162]	[0.162]	[0.164]	[0.200]	[0.172]	[0.164]
Treatment on the Treated						
ToT	**-0.0619**	**-0.0707**	**-0.0854**	**-0.0667**	**-0.0272**	**-0.0739**
	[0.124]	[0.115]	[0.126]	[0.155]	[0.123]	[0.108]
Spring Quarter	0.0067	0.0269	0.0338	0.0285	0.0292	0.0292
	[0.140]	[0.141]	[0.144]	[0.183]	[0.149]	[0.143]
ToT*Spring Quarter	**0.0461**	**0.0374**	**0.0355**	**0.0653**	**0.0401**	**0.0419**
	[0.172]	[0.173]	[0.175]	[0.226]	[0.181]	[0.173]
Control variables	No	Yes	Yes	Yes	Yes	Yes
Field of concentration dummies	No	No	Yes	No	No	No
Course dummies	No	No	No	Yes	No	No
Nationality dummies	No	No	No	No	Yes	No
Average TA evaluation	No	No	No	No	No	Yes

Robust standard errors in brackets and clustered by TA.
*** p<0.01, ** p<0.05, * p<0.1

Note: Results from a linear probability model in which the dependent variable that takes the value of 1 if the TA obtained at least one grade below seven. The variables of interest are ITT and ToT, indicator variables of the intent to treat and of the treatment on the treated, respectively. And the interaction terms of the spring quarter with the ITT and the ToT, respectively. Column 1 shows the difference of means, column 2 to 6 add the controls - age, male, English native speaker, masters degree before the PhD, number of quarters taught, a variable indicating if the TA has taught the same course before, and an indicator variable for having met the TA coordinator. Column 3 controls for field fixed effects, column 4 for course fixed effects, column 5 for nationality fixed effects and, finally, column 6 adds the mean TA evaluation. Robust errors cluster by TA.

Table 13: Attrition: Sample Size by Quarter

Variable	Observations	% assigned to treatment	Attrition rate
Sample randomized (ITT sample) in Fall 2012	55	0.582	
Sample randomized (ToT sample) in Fall 2012	48	0.521	
ITT sample: TAs teaching in Winter 2013	51	0.608	0.073
ToT sample: TAs teaching in Winter 2013	45	0.556	0.063
ITT sample: TAs teaching in Spring 2013	48	0.583	0.127
ToT sample: TAs teaching in Spring 2013	43	0.535	0.104

Note: The table shows the number of TAs - who participated in the program during Fall 2012 - that continued teaching during Winter 2013 and Spring 2013, and the corresponding attrition rate.

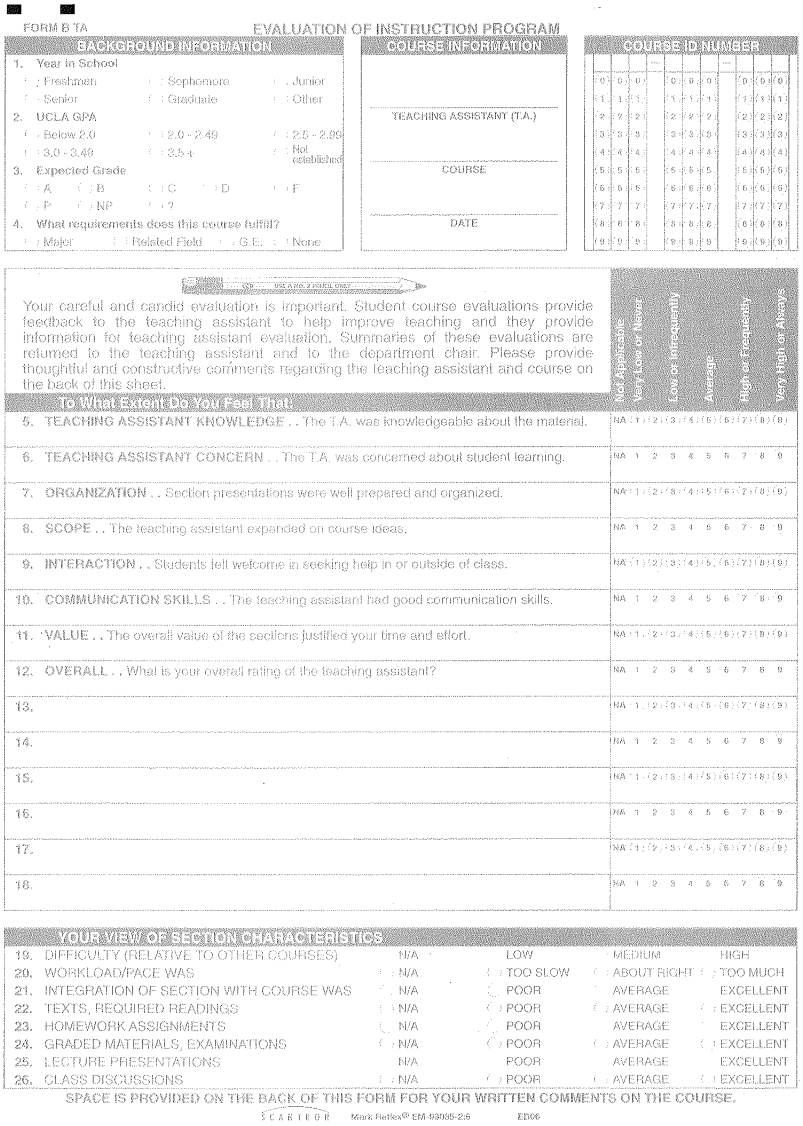
Table 14: Regression Analysis of Attrition for Winter and Spring 2013

	Winter 2013 Independent Variable: Attrition			Spring 2013 Independent Variable: Attrition		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment indicator	-0.0992		-0.0910	-0.0054		-0.0108
	[0.071]		[0.070]	[0.093]		[0.091]
Age		-0.0118	-0.0080		-0.0144	-0.0140
		[0.019]	[0.019]		[0.024]	[0.025]
I(male)		-0.0435	-0.0374		-0.0109	-0.0102
		[0.079]	[0.079]		[0.102]	[0.104]
I(English native)		-0.0816	-0.0626		-0.2163*	-0.2144*
		[0.086]	[0.087]		[0.112]	[0.114]
PhD year		0.3055***	0.2929***		0.2583**	0.2572**
		[0.089]	[0.088]		[0.113]	[0.115]
I(MA)		0.0287	0.0011		-0.0918	-0.0949
		[0.089]	[0.091]		[0.113]	[0.118]
Quarters taught		-0.0442**	-0.0424**		-0.0521**	-0.0520**
		[0.017]	[0.016]		[0.021]	[0.021]
I(taught this course before)		0.0374	0.0389		0.1057	0.1061
		[0.079]	[0.078]		[0.102]	[0.103]
I(coordinator)		-0.0861	-0.0782		-0.2199**	-0.2199**
		[0.074]	[0.074]		[0.090]	[0.091]
Constant	0.1304**	-0.1941	-0.2196	0.1304*	0.2271	0.2248
	[0.054]	[0.472]	[0.469]	[0.071]	[0.611]	[0.618]
Observations	55	55	55	55	55	55
R-squared	0.035	0.258	0.285	0.000	0.249	0.249

Note: Results from a linear probability model in which the dependent variable that takes the value of 1 if the TA did not teach during Winter (Spring) 2013. Our goal is to summarize the characteristics of the TAs that stopped teaching after the intervention.

Appendix

Table 15: Copy of the TAs Evaluation



FORM B TA

EVALUATION OF INSTRUCTION PROGRAM

BACKGROUND INFORMATION

1. Year in School
   - Freshman
   - Sophomore
   - Junior
   - Senior
   - Graduate
   - Other
2. UCLA GPA
   - Below 2.0
   - 2.0 - 2.49
   - 2.5 - 2.99
   - 3.0 - 3.49
   - 3.5 +
   - Not established
3. Expected Grade
   - A
   - B
   - C
   - D
   - F
   - P
   - NP
   - ?
4. What requirements does this course fulfill?
   - Major
   - Related Field
   - G.E.
   - None

COURSE INFORMATION

TEACHING ASSISTANT (T.A.)

COURSE

DATE

COURSE ID NUMBER

USE A NO. 2 PENCIL ONLY

Your careful and candid evaluation is important. Student course evaluations provide feedback to the teaching assistant to help improve teaching and they provide information for teaching assistant evaluation. Summaries of these evaluations are returned to the teaching assistant and to the department chair. Please provide thoughtful and constructive comments regarding the teaching assistant and course on the back of this sheet.

To What Extent Do You Feel That:	Not Applicable	Very Low or Never	Low or Infrequently	Average	High or Frequently	Very High or Always
5. TEACHING ASSISTANT KNOWLEDGE . . The T.A. was knowledgeable about the material.	NA	1	2	3	4	5	6	7	8	9
6. TEACHING ASSISTANT CONCERN . . The T.A. was concerned about student learning.	NA	1	2	3	4	5	6	7	8	9
7. ORGANIZATION . . Section presentations were well prepared and organized.	NA	1	2	3	4	5	6	7	8	9
8. SCOPE . . The teaching assistant expanded on course ideas.	NA	1	2	3	4	5	6	7	8	9
9. INTERACTION . . Students felt welcome in seeking help in or outside of class.	NA	1	2	3	4	5	6	7	8	9
10. COMMUNICATION SKILLS . . The teaching assistant had good communication skills.	NA	1	2	3	4	5	6	7	8	9
11. VALUE . . The overall value of the sections justified your time and effort.	NA	1	2	3	4	5	6	7	8	9
12. OVERALL . . What is your overall rating of the teaching assistant?	NA	1	2	3	4	5	6	7	8	9
13.	NA	1	2	3	4	5	6	7	8	9
14.	NA	1	2	3	4	5	6	7	8	9
15.	NA	1	2	3	4	5	6	7	8	9
16.	NA	1	2	3	4	5	6	7	8	9
17.	NA	1	2	3	4	5	6	7	8	9
18.	NA	1	2	3	4	5	6	7	8	9

YOUR VIEW OF SECTION CHARACTERISTICS

19. DIFFICULTY (RELATIVE TO OTHER COURSES)	N/A	LOW	MEDIUM	HIGH
20. WORKLOAD/PACE WAS	N/A	TOO SLOW	ABOUT RIGHT	TOO MUCH
21. INTEGRATION OF SECTION WITH COURSE WAS	N/A	POOR	AVERAGE	EXCELLENT
22. TEXTS, REQUIRED READINGS	N/A	POOR	AVERAGE	EXCELLENT
23. HOMEWORK ASSIGNMENTS	N/A	POOR	AVERAGE	EXCELLENT
24. GRADED MATERIALS, EXAMINATIONS	N/A	POOR	AVERAGE	EXCELLENT
25. LECTURE PRESENTATIONS	N/A	POOR	AVERAGE	EXCELLENT
26. CLASS DISCUSSIONS	N/A	POOR	AVERAGE	EXCELLENT

SPACE IS PROVIDED ON THE BACK OF THIS FORM FOR YOUR WRITTEN COMMENTS ON THE COURSE.

SCANTRON Mark Reflex® EM-93095-2.5 ED06

Note: This is a copy of the actual evaluation that students fill out at the end of the quarter (usually in the last two weeks of the quarter). The normal procedure is that the TA hands the evaluations to the students and leaves the room after instructing one of the students to take them in a sealed envelope to the Department's Office. This way, the TA has no chance to change or modify the scores of the students.

7 References

Allen, Liz (2002). "Consenting Adults in Private - Union and Management Perspectives on Peer Observation of Teaching." Working Paper of the Higher Education Academy.

Atwood, C. H., Taylor, J. W., & Hutchings, P. A. (2000). "Why are chemists and other scientists afraid of the peer review of teaching?" Journal of Chemical Education, 77: 239–244.

Allgood, Sam, William B. Walstad, and John J. Siegfried. (2015). "Research on Teaching Economics to Undergraduates," Journal of Economic Literature, 53(2): 285–325.

Braga, Michaela, Marco Paccagnella and Michelle Pellizari. 2014. "Evaluating students' evaluations of professors," Economics of Education Review, 41: 71-88.

Becker, William E., and Michael Watts. 1999. "How Departments of Economics Evaluate Teaching." American Economic Review 89 (2): 344–49.

Beleche, Trinidad, David Fairris and Mindy Marks. 2012. "Do course evaluations truly reflect student learning? Evidence from an objectively graded post-test," Economics of Education Review, 31(5): 709-719.

Bell, Amani and Mladenovic, Rosina. (2008). "The Benefits of Peer Observation of Teaching for Tutor Development." High Education.

Bell, Amani. (2005). "Peer Observation and Teaching in Australia". Report of the Higher Education Academy.

Bettinger, Eric and Bridget Long. (2011) "Do College Instructors Matter? The Effects of Adjuncts on Students' Interests and Success." Review of Economics and Statistics.

Borjas, George J. (2000). "Foreign-Born Teaching Assistants and the Academic Performance of Undergraduates," The American Economic Review.

Bureau of Labor Statistics. 2002. Occupational Employment and Wages.

Bureau of Labor Statistics. 2014. Occupational Employment and Wages.

Carrell, Scott E. and James E. West. (2010) "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors." Journal of Political Economy.

Carroll, Gregory. (1980). "Effects of Training Programs for University Teaching Assistants: A Review of Empirical Research," The Journal of Higher Education.

Cheng, Jacqueline and Herbert Marsh. 2010. "National Student Survey: are differences between universities and courses reliable and meaningful?" Oxford Review of Education, 36(6): 693-712.

Croteau, D. & Hoynes, W. (1991) A transitional teaching experience: learning groups and the first-time teacher, Teaching Sociology, 19(1): 28–33.

Cunha, Jesse M., and Trey Miller. (2014). "Measuring value-added in higher education: Possibilities and limitations in the use of administrative data." Economics of Education Review, 42: 64-77.

Dalgaard, K. A. (1982). Some effects of training on teaching effectiveness of untrained university teaching assistants. Research in Higher Education, 17(1): 39–50.

Danielson, Charlotte. (2011). "The Framework for Teaching Instrument," The Danielson Group.

David, Adonis and Macayan, Jonathan. (2010). "Assessment of Teacher Performance," The Assessment Handbook Vol. 3.

Dye, Ronald. (1986). "Optimal Monitoring Policies in Agencies," The RAND Journal of Economics, RAND Corporation.

Ehrenberg, Ronald G. and Liang Zhang. (2005) "Do Tenure and Tenure-Track Faculty Matter?" Journal of Human Resources.

Gosling, D. (2005), Peer observation of teaching: SEDA Paper 118 (Birmingham: SEDA).

Gosling, D., &O'Connor, K. M. (2006). From peer observation of teaching to reviewofprofessional practice (RPP): A model for continuing professional development. Educational Developments, 7(3): 1–6.

Gosling, D., & O'Connor, K. M. (Eds.). (2009). Beyond peer observation of teaching. London: Staff and educational development association (SEDA).

Gilbert, C. P., & McArthur, J. F. (1975). In-service teacher preparation of French Graduate Assistants: design and evaluation. The French Review, 48(3), 508–521.

Hammersley-Fletcher, L., & Orsmond, P. (2004). Evaluating our peers: Is peer observation a meaningful process? Studies in Higher Education, 29, 489–503.

Hanushek, Eric. 1986. "The Economics of Schooling: Production and Efficiency in Public Schools." Journal of Economic Literature 24(3): 1141–77.

Hanushek, Eric. 1995. "Interpreting Recent Research on Schooling in Developing Countries." World Bank Research Observer 10(2): 227–46.

Hanushek, Eric. (2007). "Education Production Functions". Palgrave Encyclopedia.

Hoffman, Florian and Philip Oreopoulos. (2009) "Professor Qualities and Student Achievement". Review of Economics and Statistics.

Holmstrom, Bengt. (1979). "Moral Hazard and Observability". The Bell Journal of Economics, Rand Corporation.

Ingersoll, R. 2003. Out-of-Field Teaching and the Limits of Teacher Policy. Seattle, Wash.: Center for the Study of Teaching and Policy.

Kanagaretnam, Kiridaran, Robert Mathieu, and Alex Thevaranjan. 2003. "An Economic Analysis of the Use of Student Evaluations: Implications for Universities." Managerial and Decision Economics 24 (1): 1–13.

Kell, C., & Annetts, S. (2009). Peer review of teaching embedded practice or policyholding complacency? Innovations in Education and Teaching International, 46, 61–70.

Kherfi, Samer. 2011. "Whose Opinion Is It Anyway? Determinants of Participation in Student Evaluation of Teaching." Journal of Economic Education 42 (1): 19–30.

Koedel, C. and Beets, J. (2007) "Re-Examining the Role of Teacher Quality In the Educational Production Function". University of Missouri Working Paper.

Koedel, Cory and Betts, Julian (2008). "Test-Score Ceiling Effects and Value-Added Measures of School Quality". JSM Proceedings, Social Statistics Section.

Lavy, Victor. 2002. "Evaluating the Effect of Teachers' Group Performance Incentives on Pupil Achievement." Journal of Political Economy 110(6): 1286–317.

Lavy, Victor. 2007. "Using Performance-Based Pay to Improve the Quality of Teachers", The Future of Children, Spring: 87-110.

Lavy, Victor. 2009. "Performance Pay and Teachers' Effort, Productivity and Grading Ethics", American Economic Review 99(5): 1979-2011.

Lawrenz, F. et al. (1992) Training the teaching assistant, Journal of College Science Teaching, 22(2), 106–109.

Laat, Joost de. (2005). "Moral Hazard and Costly Monitoring: The Case of Split Migrants in Kenya". Mimeo.

Lenton, Pamela. 2015. "Determining student satisfaction: An economic analysis of the National Student Survey," Economics of Education Review, 47:118–127.

Levitt, Steven and Jacob, Brian. (2003). "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." Quarterly Journal of Economics.

List, John and Imran, Rasul. (2011). "Field Experiments in Labor Economics" Handbook of Labor Economics V. 4a. Chapter 2.

Lomas, L., & Nicholls, G. (2005). Enhancing teaching quality through peer review of teaching. Quality in Higher Education, 11, 137–149

Marsh, H. W. (1983). "Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics". Journal of Educational Psychology.

McCoy, James P., and Martin I. Milkman. 2010. "Do Recent PhD Economists Feel Prepared to Teach Economics?" Journal of Economic Education 41 (2): 211–15.

McGoldrick, KimMarie, Gail Hoyt, and David Colander. 2010. "The Professional Development of Graduate Students for Teaching Activities: The Students' Perspective." Journal of Economic Education 41 (2): 194–201

McPherson, Michael A., R. Todd Jewell, and Myungsup Kim. 2009. "What Determines Student Evaluation Scores? A Random Effects Analysis of Undergraduate Economics Classes." Eastern Economic Journal 35 (1): 37–51.

National Center for Education Statistics. 2001. Institutional Policies and Practices: Results from the 1999 National Study of Postsecondary Faculty, Institutional Survey.

National Center for Education Statistics. 2000. 1999-2000 National Postsecondary Student Aid Study.

National Center for Education Statistics. 2013. Digest of Education Statistics.

Norris, Timothy. (1991). "Nonnative English-Speaking Teaching Assistants and Students Performance". Research in Higher Education.

Numberg, Peter, Morton Shapiro and David Zimmerman. 2012. "Students choosing colleges: Understanding the matriculation decision at a highly selective private institution," Economics of Education Review, 31(1): 1-8.

Park, Chris. (2004). "The graduate teaching assistants (GTAs): lessons from North America". Teaching in Higher Education.

Perdendergast, Canice. (1999). "The Provision of Incentives in Firms". Journal of Economic Literature.

Quinlan, K. and G. Akerlind. (2000). "Factors Affecting Departmental Peer Collaboration for Faculty Development: Two Cases in Context". Higher Education Vol. 40.

Rice, Jennifer. 2003. Teacher Quality: Understanding the Effectiveness of Teacher Attributes. Washington, D.C.: Economic Policy Institute.

Rivkin, S. G., E. A. Hanushek, and J. F. Kain. 2005. "Teachers, Schools, and Academic Achievement." Econometrica 73(2): 417-458.

Robinson, J. B. (2000) New teaching assistants facilitate active learning in Chemistry laboratories: promoting teaching assistant learning through formative assessment and peer review, Journal of Graduate Teaching Assistant Development, 7(3), 147–162.

Sachs, Judyth and Mitch Parsell. 2014. "Professional Learning and Development in Schools and Higher Education." New York: Springer.

Salemi, Michael K., and William B. Walstad. 2010. Teaching Innovations in Economics: Strategies and Applications for Interactive Instruction. Cheltenham, U.K. and Northampton, Mass.: Elgar.

Sanders, W., and J. Rivers. 1996. Cumulative and Residual Effects of Teachers on Future Student Academic Achievement. Knoxville: University of Tennessee Value-Added Research and Assessment Center.

Siegfried, John J., and William B. Walstad. 1998. "Research on Teaching College Economics." In Teaching Undergraduate Economics: A Handbook for Instructors, edited by William B. Walstad and Phillip Saunders, 141–66. New York: Irwin/McGraw-Hill.

Sparks, G.M. (1986). "The effectiveness of alternative training activities in changing teaching practices". American Educational Research Journal.

Thomas, Susan, Mathew Abraham, Qiu Ting Chie, Sony Jalarajan Raj and Loo-See Beh. 2014. "A Qualitative Review of Literature on Peer Review of Teaching in Higher Education: An Application of the SWOT Framework." Review of Educational Research 84(1): 112-159.

Tuckman, Howard P. (1975). "Teacher Effectiveness and Student Performance". The Journal of Economic Education.

Umansky, Ileana and Vargas, Emiliana. (2005). "Incentives to Improve Teaching: Lessons from Latin America". World Bank Publications.

Walstad, William B., and William E. Becker. 2010. "Preparing Graduate Students in Economics for Teaching: Survey Findings and Recommendations." Journal of Economic Education 41 (2): 202–10.

Watts, Michael and Gerald J. Lynch. (1989). "The principles courses revised". The American Economic Review, Papers and Proceedings.

Webber, Douglas A., and Ronald G. Ehrenberg. (2010). "Do expenditures other than instructional expenditures affect graduation and persistence rates in American higher education?" Economic of Education Review, 29(6): 947–958.