Buurman, Margaretha; Delfgaauw, Josse; Dur, Robert; Zoutenbier, Robin

Working Paper

When Do Teachers Respond to Student Feedback? Evidence from a Field Experiment

IZA Discussion Papers, No. 12907

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Buurman, Margaretha; Delfgaauw, Josse; Dur, Robert; Zoutenbier, Robin (2020) : When Do Teachers Respond to Student Feedback? Evidence from a Field Experiment, IZA Discussion Papers, No. 12907, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/215303

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12907

When Do Teachers Respond to Student Feedback? Evidence from a Field Experiment

Margaretha Buurman
Josse Delfgaauw
Robert Dur
Robin Zoutenbier

JANUARY 2020

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12907

When Do Teachers Respond to Student Feedback? Evidence from a Field Experiment

Margaretha Buurman
Free University Amsterdam

Josse Delfgaauw
Erasmus University Rotterdam and Tinbergen Institute

Robert Dur
Erasmus University Rotterdam, Tinbergen Institute, CESifo and IZA

Robin Zoutenbier
Ministry of Finance

JANUARY 2020

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

When Do Teachers Respond to Student Feedback? Evidence from a Field Experiment*

We ran a field experiment at a large Dutch school for intermediate vocational education to examine whether the response of teachers to student feedback depends on the content of the feedback. Students evaluated all teachers, but only a randomly selected group of teachers received feedback. Additionally, we asked all teachers before as well as a year after the experiment to assess their own performance on the same items. We find a precisely estimated zero average treatment effect of receiving student feedback on student evaluation scores a year later. However, teachers whose self-assessment before the experiment is much more positive than their students' evaluations do improve significantly in response to receiving feedback. We also find that provision of feedback reduces the gap between teachers' self-assessment and students' assessment, but only to a limited extent. All of these results are driven by the female teachers in our sample; male teachers appear to be unresponsive to student feedback.

JEL Classification: C93, I2, M5

Keywords: field experiment, feedback, teachers, student evaluations, self-assessment, gender differences

Corresponding author:
Robert Dur
Erasmus University Rotterdam
Department of Economics H9-15
P.O. Box 1738
3000 DR Rotterdam
The Netherlands

E-mail: dur@ese.eur.nl

* We gratefully acknowledge comments and suggestions by the co-editor and three anonymous reviewers of this journal, Karen van der Wiel and Marc van der Steeg, seminar participants at Goethe University Frankfurt, the University of Hamburg, the University of Konstanz, and the Netherlands Bureau for Economic Policy Analysis, and conference participants at the 2014 Work- shop in Public Organizations at the University of Vienna, the 2015 Amsterdam Symposium on Behavioral and Experimental Economics, the 2015 Workshop Effectiveness of Interventions in Education at the Dutch Ministry of Economic Affairs, the 2016 Workshop on Pro-social Motivation at Work in Rotterdam, and the 2017 Workshop on Recognition and Feedback in Rotterdam. An earlier version of this paper circulated under the title .The Effects of Student Feedback to Teachers: Evidence from a Field Experiment. The data used in this study are proprietary. Researchers interested in replicating our findings can access the data in our presence at the Erasmus University Rotterdam.

1 Introduction

Regular provision of feedback to employees is common practice in many organizations. Feedback often serves as a means to provide recognition to good performers as well as to help employees learn about how to improve one's performance. Several recent studies, conducted in a variety of organizations and contexts, have shown that the provision of feedback can have sizeable positive effects on performance (Azmat and Iriberri 2010, 2016, Blanes i Vidal and Nossol 2011, Kuhnen and Tymula 2012, Tran and Zeckhauser 2012, Delfgaauw et al. 2013, Gerhards and Siemer 2016, Azmat et al. 2019). On the contrary, Barankay (2012) and Bandiera et al. (2013) find an adverse effect of feedback on performance.

Providing employees with feedback has also become increasingly prevalent in education. Many schools use students' evaluations of teachers to enable and motivate teachers to improve teaching.[1] Moreover, students' evaluations sometimes play a role in tenure, bonus, and promotion decisions (Watts and Becker, 1999). There is by now an extensive literature that studies the use of students' evaluations in education. Cohen (1980) and Marsh (2007) present overviews of the literature.

This paper addresses the question under what conditions teachers are responsive to student feedback. We are particularly interested in how teachers' responses depend on the content of the feedback. With the exception of Centra (1973), this question has not received any attention in the empirical literature so far.

We study a large group of teachers who work at a school that has so far not made use of any formal student feedback system. We collect student evaluations on all teachers as well as teachers' self-assessments. We hypothesize that if a teacher receives student evaluation scores that are very similar to his or her self-assessment, (s)he has little reason to adjust teaching practices, and so (s)he is likely to receive similar student evaluations a year later. In contrast, if students express views that are much less favorable than the teacher's self-assessment, the teacher may try to improve, resulting in better scores a year later. Teachers who receive student evaluations that are more positive than their self-assessment may slack down, as they may infer that less effort is needed in order to be evaluated well.[2]

[1] Moreover, in some schools teachers are evaluated by external experts or by peers, see Taylor and Tyler (2012), Briole and Maurin (2019), and Burgess et al. (2019).

[2] The underlying assumptions here are that teachers care about their student evaluation

To test these hypotheses, we set up a field experiment at a large Dutch school for intermediate vocational education. Student evaluations were introduced for all teachers in the form of an electronic questionnaire consisting of 19 items. We implemented a feedback treatment where a randomly chosen group of teachers received the outcomes of their students' evaluations. The other group of teachers was evaluated as well but did not receive any personal feedback. We examine the effect of receiving feedback on student evaluations a year later performed by a new group of students. Our key research objective is to find out whether the effect of feedback depends on how student evaluations differ from the teacher's own performance assessment on the same items. For that purpose, we collect data on teachers' self-assessed performance both before and a year after the experiment.

The results of our experiment show that receiving feedback has on average no effect on feedback scores of teachers a year later. We find a precisely estimated zero average treatment effect of 0.04 on a 5-point scale with a standard error of 0.05. Our result differs somewhat from the findings of the existing studies mentioned above. A possible explanation for the lack of an average treatment effect in our study may be that we investigate the effect of feedback in the long run. Feedback may affect short-run performance, but the effect may fade away in the long run, as Azmat et al. (2019) show in the context of providing relative performance information to students. While earlier studies on student feedback commonly consider the effects of receiving feedback within a semester, we study the effect of feedback on student evaluations a full year later.

Regarding the content of the feedback, we find – in line with our predictions – that there is no effect of the feedback treatment for teachers who evaluate themselves similarly to the students' evaluation. The estimate of the treatment effect for these teachers is very close to zero. We do find a significant positive treatment effect for teachers who learn that their own assessment is much more favorable than their students' evaluation.

Our findings are well in line with Centra (1973), the only prior study – to our knowledge – investigating whether teachers' response to student evaluations depends on the discrepancy between teachers' self-assessment and their students' evaluations. Among a sample of about 350 teachers at 5 different colleges in the US, he finds on average little effect of mid-semester feedback on end-of-semester student ratings. However, among teachers for

scores, that these scores are affected by teachers' effort and talent, that teachers may not be perfectly informed about their talent, and that talent and effort are substitutes. If talent and effort are highly complementary, the effort responses may be opposite (positive in response to learning that students are more positive than the self-assessment and negative when students are less positive). See also the model in Azmat et al. (2019).

whom students' mid-semester ratings fell short of their own assessment, end-of-semester ratings increased more strongly as compared to similar teachers who did not receive feedback. Our study finds, in a different population, similar results that hold over the period of a full year.

How a teacher's student evaluation scores compare to the student evaluation scores of her colleagues may also matter for the effect of receiving feedback. In our experiment, all teachers –both in treatment and control– were informed about the average of the evaluation outcomes of the teachers in their team. This implies that some teachers in the treatment group learn that they perform better than their direct colleagues, while others learn that they perform worse. Relative performance information may matter for the performance of teachers when they care about their status (Moldovanu et al. 2007, Besley and Ghatak 2008, Auriol and Renault 2008) or when teachers want to conform to social norms (Bernheim 1994, Sliwka 2007). Our results show that the treatment effect is very close to zero for teachers who received student evaluation scores that are better than those of their teammates. We do find a positive but small (and only marginally significant) effect of feedback for workers who have worse student evaluation scores than their team on average.

An additional response of teachers to receiving student evaluations that conflict with their self-assessment is to adjust their self-assessment. We find only small effects of the feedback treatment on the self-assessment of teachers. Teachers who learn that their students' evaluations are on average better than their self-assessment do not update their self-assessment. Teachers who learn that their students' evaluations are worse than their self-assessment do lower their self-assessment of performance, but only to a limited extent.

When we presented these findings in seminars and conferences, we were often asked whether there are gender differences in the response to feedback. The literature suggests a couple of reasons for why a gender difference in responses may arise. Roberts and Nolen-Hoeksema (1994) and Johnson and Helgeson (2002) find that women are more likely to internalize feedback than men, in particular when the feedback is negative. In lab experiments, Mobius et al. (2007) and Buser et al. (2018) find gender differences in updating in response to relative performance, where women turn out to be more conservative in updating after receiving relative performance feedback than men. Azmat and Iriberri (2016) find that males' performance improves significantly more than females' performance after receiving relative performance feedback (in addition to feedback on individual performance). This gender difference does not depend on the content of feedback, and is stronger under individual pay-for-performance than under flat wages.

Performing our analysis separately for male and female teachers, we find

that the pattern of responses as described above is entirely driven by female teachers. Whereas male teachers hardly respond to feedback independent of the content, we find that female teachers' student evaluation scores increase significantly after learning that their student evaluation score falls below their self-assessment score as well as when they learn their score is worse than that of their team. Moreover, in contrast to male teachers, female teachers adjust their self-assessment downwards after learning that students rate them less favorably than they rated themselves. As this is an ex post analysis, these results should be considered as exploratory. Further research on gender differences in response to student feedback is warranted.

Finally, we investigate whether receiving feedback and the content of the feedback have an effect on teachers' job satisfaction. Receiving information about performance might affect teachers' job satisfaction when teachers intrinsically care about their performance (as in e.g. Besley and Ghatak 2005 and Delfgaauw and Dur 2008) or when they enjoy being perceived as a competent or dedicated teacher (as in Suurmond et al. 2004 or Benabou and Tirole 2006). In either case we would expect that job satisfaction of teachers in the treatment group increases with the difference between student feedback and teacher's self-assessment. Earlier work by Ryan et al. (1980) shows that the introduction of student evaluations negatively affects job satisfaction on average. Our results show that providing teachers with feedback on their performance has no significant effect on their job satisfaction. We also find that the effect does not depend on the content of feedback.

An important limitation of our study is that we have no data on standardized student test scores or other objective measures of student performance. Hence, we cannot examine whether providing feedback affects students' performance and/or teachers' value added. Carrell and West (2010) and Braga et al. (2014) present evidence that student evaluation scores are *negatively* correlated with teachers' value-added, raising doubts about the usefulness of student evaluations. Beleche et al. (2012), on the other hand, find a robust positive association between student learning and course evaluations. Likewise, Mengel et al. (2019) find a positive correlation for male teachers, while there is little correlation for female teachers in their sample. While these results have raised concerns about the meaning and usefulness of student evaluations in general, it is unclear what they exactly mean in our context. First, student evaluations have been rarely studied in the context of vocational education. The mixed findings on the correlation between student evaluation scores and student learning all stem from a tertiary education context (colleges and universities). It is unclear what to conclude from these studies for vocational education. Second, our student evaluation form goes way beyond the typical form used in tertiary education with only a handful

of simple questions. Instead, the school took great care to design a rather extensive set of 19 questions which fit the local situation well and which the school's management and representatives of the teachers believe might help the teachers to improve themselves. Third, even if in our context the student evaluation scores would have no (or a negative) correlation with learning outcomes, they may bear relevance in another way. The answers to the 19 questions in the student evaluation we study together sketch a picture of student satisfaction with the teachers, which may be of independent importance for students' decision to continue their current study as well as students' intentions to do a follow-up study. Fourth, and this holds more generally, a correlation between student evaluation scores and objective learning outcomes (be it positive, negative, or zero) is not necessarily a good predictor of what an improvement in student evaluations caused by teachers' response to students' feedback means for objective learning outcomes, Even if student evaluation scores and learning outcomes are negatively correlated (as is the case in some of the studies mentioned above), this need not be a causal effect, and an improvement in student evaluation scores caused by an intervention might go hand in hand with an improvement in student learning outcomes. Unfortunately, we are not able to shed light on these issues in our context, as we lack objectives measures of student performance.

We proceed as follows. The next section provides a detailed description of the field experiment. Section 3 reports the descriptive statistics of the sample. In section 4 we describe our empirical strategy. The results of the field experiment are presented in section 5. We discuss gender differences in response to feedback in section 6. Finally, section 7 concludes.

2 Experimental design

2.1 Background

The field experiment took place at a Dutch school for intermediate vocational education between September 2011 and February 2013. The school offers education to teenagers (usually in the age range from 16 to 20) and (young) adults. The offered curricula prepare for a large number of occupations, including technical professions, administrative jobs, maritime professions, and jobs in information technology, health care, and the hospitality sector. In all fields, there are multiple programs that differ by level and duration. The durations of programs vary between one and four years.

All teachers are assigned to teams that are supervised by a manager. The teams are organized around educational fields. Each team consists of

roughly 10 to 20 teachers. Teachers teach one or several courses to a number of different classes of students. Teachers of general subjects (such as language or math) typically teach in multiple fields, while most teachers of field-specific courses (such as cooking or inland shipping) only teach students within their own field. Depending on the field of education, the average class size is 10 to 30 students. Students can have the same teacher for different courses in their program.

In 2011, the school had almost 8,000 students and about 470 teachers divided over 27 teams. The school merged in 2012 with another intermediate vocational education school, which increased the number of students to about 9,500 and the number of teachers to about 550. In 2013, the school had 9,000 students and 520 teachers. The merger was officially announced in March 2012 and formally took place on August 1, 2012. The merger did not interfere with our experiment in that the organizational structure as well as the composition of the teams in the experiment remained largely unchanged. However, the merger did result in a higher attrition of teachers, which we shall analyze in depth in the next section.

The teachers in the experiment had not received individual feedback from student evaluations at this school in the past. During the experiment, no other individual feedback based on student evaluations was provided to the teachers. The school does participate in a national survey on student satisfaction, which provides information about the student evaluations of the school and of educational fields. Furthermore, most teachers have annual performance interviews with their manager. Finally, in 2011 teachers participated in a 360 degree evaluation, which included feedback from their manager, colleagues, and external clients (such as companies that provide internships), but not from students. None of these alternative types of feedback differed between teachers in the treatment group and the control group in our experiment.

Teachers at this school earn a flat wage. The school originally intended to follow up on this feedback experiment with another, government-funded experiment aimed at testing the effects of individual incentive pay for teachers, partially based on student evaluation scores. However, this plan was abandoned in May 2012 due to central government budget cuts. The school did continue the yearly student evaluations after the experiment ended.

2.2 The questionnaire

In the year prior to our collaboration with the school, six teams had implemented student evaluation surveys as a pilot. This 19-item survey was designed to give teachers valuable feedback that might help them improve

their teaching. It had the support of teacher representatives and the school's management. After analyzing the outcomes of this pilot survey, we agreed to use the same questions in our study, with some minor adjustments to the wording. The six pilot-teams are not part of our experiment, which took place within the remaining 21 teams. The final version of the questionnaire can be found in the Appendix. It consists of 19 statements, to which students could respond on a 5-point scale ranging from 'disagree' to 'agree',[3] as well as a space for comments and recommendations.

The questionnaire items can be grouped into 4 categories: didactical skills (items 1 - 6), pedagogical skills (items 7 - 11), organizational aspects (items 12 - 15), and interpersonal skills (items (16 - 19). The school's management and the teachers considered these to be the most relevant aspects of teaching at the school.[4]

The completion of the surveys by students took place during class hours, under the supervision of (preferably) a person who was not evaluated by that class of students. Students went to a separate classroom, where each of them had access to a computer to complete the surveys. It was decided that students would evaluate a maximum of three teachers. Asking students to evaluate more teachers was deemed undesirable, as students might lose interest after filling out several questionnaires. The team managers decided which teachers would be evaluated by a particular class of students. In the data, the number of teachers evaluated by a student ranges from 1 to 5. Nearly all teachers in the 21 teams were evaluated by students.[5]

[3]In addition, students could respond "Do not know / not applicable" to a statement. Throughout the analysis, we treat such responses as missing observations. Alternatively, we could drop questionnaires with partial non-response altogether. This reduces the sample size to quite some extent, but does not affect any of our main conclusions.

[4]The four aspects of teaching addressed in the survey (didactical skills, pedagogical skills, organizational aspects, and interpersonal skills) have strong within-category correlations. Cronbach's Alpha for these categories in the student evaluations in 2011 are 0.97, 0.93, 0.85, and 0.90, respectively. For the teachers' self-assessments, the Cronbach's Alpha's are weaker: 0.65, 0.65. 0.45, and 0.40, respectively. Factor analyses on the student evaluation scores and the teachers' self-assessments return one factor that loads positively on all 19 items, albeit more strongly so for the student evaluations than for the self-assessments. In both sets of questions, the remaining factors explain only a limited amount of the variation and do not have a clear interpretation.

[5]We lack the data needed to compare the characteristics of students and classes that did perform evaluations with those that did not and, hence, cannot gauge how team managers selected the classes that performed evaluations. Team managers did not have formal incentives linked to these evaluations. Moreover, as discussed in Section 2.3, any selection by team managers in the first year of evaluations would be orthogonal to the teachers' assignment to treatment and control, as we performed the assignment after the student evaluations were completed.

2.3 Set-up of the experiment

The experiment is based on two waves of student evaluations of teachers. In both 2011 and 2012, students were asked to evaluate the performance of teachers during the first teaching period of the school year. This period runs from September to mid-November. The questionnaires were administered between mid-November and mid-December. In both years, in the same period as the student evaluations took place, all teachers were asked to complete a self-assessment questionnaire on the same items as contained in the student evaluation questionnaire.[6]

Before the start of the school year in 2011, teachers were informed through an information bulletin that student evaluations would take place. The information bulletin also stated that a random half of the teachers would receive their evaluation scores, so as to enable an evaluation of the effects of feedback provision. Exactly which teachers would receive their scores was determined after the student evaluations and teacher self-assessments had taken place, through a randomization procedure described below. Before the start of the school year in 2012, teachers were informed that another round of student evaluations would take place, and that all teachers would receive their scores this time. Our experiment thus yields an estimate of the effect of feedback provision on subsequent performance. Our design does not enable us to assess the effect of the anticipation of feedback provision (as all teachers anticipated that they might receive feedback), nor can we assess the possible effects of performance measurement (because all teachers knew that their performance would be measured).

After the first wave of evaluations had taken place, we randomly assigned teachers to treatment and control. Within each team, we stratified the assignment by average student evaluation score and by the difference between teachers' average self-assessment score and average student evaluation score, in the following way. Within each team, we ranked teachers by their average score (over all students that evaluated them) on all 19 statements except statements 14 and 15.[7] Based on this ranking, we created three equally large strata. Within these strata, we ranked all teachers based on the difference between their average self-assessment scores and their average student

[6] In contrast to the student evaluation form, the questionnaire for teachers did not contain "Do not know / not applicable" as a possible answer category. Only 5 teachers refrained from answering one or more items. We excluded these teachers from the sample.

[7] We excluded statements 14 and 15 here because these consider factual statements regarding time taking for answering e-mails and grading (see the Appendix). We expected that on these items, students' answers were unlikely to provide any new information to teachers. On the other 17 items, students' experience may differ from the teacher's perception and, hence, these are more likely to contain novel information for the teacher.

evaluation score, both based on the same 17 items. Using this ranking, we alternated the assignment of teachers to treatment and control, using a random device to determine whether the teachers in odd positions or the teachers in even positions were placed in the treatment group.[8] This procedure helps to create balance between the treatment group and the control group in terms of average student evaluation score as well as in terms of the gap between student evaluation scores and self-assessment score. This stratification increases the power of our analysis (List et al. 2011), and has the additional benefit of credibly indicating that the hypothesis on whether the effect of feedback depends on the discrepancy between self-assessment and student evaluation score was formulated ex ante.

The teachers in the treatment group received their feedback in February 2012 through e-mail. It contained the average student evaluation score on each of the 19 items, both for all evaluations together as well as split out by class. It also contained the average evaluation score over all items, again averaged over all evaluations as well as by class. Furthermore, it included the teacher's self-assessment scores, on all items as well as the overall average. Lastly, it contained the average student evaluation score of all teachers in the teacher's team, on all 19 items as well as the overall average. Note that in the team scores, the student evaluations of teachers in the control group are included. The team managers also received this feedback of the teachers in the treatment group (but not of the teachers in the control group). The teachers in the control group did not receive their individual student evaluation scores, but they did receive their self-assessment scores as well as the team scores.[9]

To study whether and when teachers respond to feedback, our main outcome measure is average student evaluations one year later. Unfortunately, there are no 'objective' performance measures available. During the period of our experiment, there were no standardized tests at this school. Moreover, as students had about half of their teachers who did and the other half of their teachers who did not receive feedback, we cannot use passing rates, drop-out rates, or grade averages as performance measures.

[8] Teachers who did not complete the self-assessment were randomly assigned to treatment and control.

[9] The e-mail with or without individual student evaluation scores was also the first moment at which a teacher learned whether he or she would receive the individual feedback or not. Possibly, teachers in the treatment and control groups discussed the feedback amongst each other after receiving the e-mails. However, as we assigned teachers to treatment and control through stratified randomization within teams based on individual student evaluation scores, teachers in the control group were unable to infer their individual student evaluation scores, even if they learned all individual evaluation scores received by teachers in their team.

Between mid-November and mid-December in 2012, we conducted the second wave of student evaluations using the same questionnaire and the same procedure. As in the previous year, students were asked to evaluate their teacher based on their experience in the first teaching period of the school year. Furthermore, all teachers were asked to complete the self-assessment questionnaire again. This allows us to study whether teachers' self-assessment responds to students' feedback. All teachers received their student evaluation scores in February 2013.

Lastly, to examine the effect of feedback on teachers' job satisfaction, we use data from an employee satisfaction survey that was conducted independently of this experiment in November 2012. We measure a teacher's job satisfaction by her answer to the statement: "I am satisfied with working at [school name]". Respondents could answer on a 5-point scale ranging from "not at all satisfied" to "fully satisfied".[10]

3 Data description

In the first wave of student evaluations, 323 teachers are evaluated. These teachers are randomly assigned to the treatment or the control group, in the manner described in the previous section. In the second wave of student evaluations, 242 out of these 323 teachers are again evaluated. Hence, 81 teachers drop out of our sample between the first and second wave of student evaluations. Our estimations are based on the remaining 242 teachers, of whom 116 teachers have been assigned to the treatment group, while the remaining 126 teachers are in the control group. Over the two waves, we have a total of 15,194 student evaluation scores of these teachers, 7,951 in 2011 and 7,243 in 2012. The number of evaluations per teacher may differ due to differences in class size or differences in response rates across classes. In each year, less than 10 teachers have fewer than 8 student evaluations, and at most 7 teachers have more than 60 student evaluations. Our final sample contains evaluations by 5,761 unique students.[11] Below, we first provide

[10] The job satisfaction question is part of the organization's employee satisfaction survey that is conducted on a yearly basis. Unfortunately, both the wording of the job satisfaction question as well as the answer scales differ between the year before and the year after we provided feedback to a random subset of the teachers. As a result, it is difficult to compare job satisfaction before receiving feedback to job satisfaction after receiving feedback.

[11] Only 1,010 students evaluate teachers in both years. Out of the 7,243 student evaluations in 2012, 776 evaluations concern a student who evaluates a teacher whom he/she had also evaluated in 2011 (this set of observations contains 623 different students and 108 different teachers). Hence, we do not have enough observations to perform a meaningful analysis with 'stable' student-teacher matches only. Neither removing these 776 evalua-

descriptive statistics for the 242 teachers in the analysis and subsequently discuss attrition.

Table 1 reports descriptive statistics for the teachers in our analysis. In the first wave, teachers are on average evaluated by about 33 students. The average evaluation score of a teacher in 2011 is 4.12 on a 5-point scale. On average, teachers' self-assessment score is 4.60, which is considerably higher than the evaluations by their students. Table A.1 in the Appendix gives average evaluation scores and self-assessment scores per questionnaire item. The table shows that students are on average quite satisfied with nearly all aspects, but that teachers rate themselves consistently higher than students. The correlations between teachers' self-assessment score and the average student evaluation score reported in column 5 of Table A.1 are quite low. As a result, there is substantial variation in the difference between a teacher's self-assessment and his/her average student evaluation score. Correlations across items, as reported in Table A.2, are considerably higher in the student evaluations than in the teachers' self-assessment.

Columns 1 and 2 in Table 1 show that the average evaluation score in 2011 hardly differs between teachers in the treatment group and teachers in the control group. The difference is 0.05 and statistically insignificant. We also find no significant difference in teachers' self-evaluations between the treatment group and the control group. On observable characteristics, teachers in the two groups are also comparable. Teachers in the treatment group are slightly less likely to be female, are a bit younger, have shorter tenure, and work less hours on average than teachers in the control group. Only the differences in working hours and tenure are marginally significant at the 10-percent level.[12]

Figure 1 shows the average student evaluation score in the treatment group and the control group for both years. For both groups, the average evaluation score in the first year is slightly higher than the average score in the second year. This reduction in evaluation scores is slightly smaller for teachers in the treatment group. Figures 2 and 3 show the distribution of the student evaluation scores in the treatment group and the control group, for the first and second year, respectively. Figures 2 shows that our stratified randomization was successful in balancing teachers' 2011 average student evaluation scores between the treatment group and the control group. The distributions of the 2012 average evaluation scores do not markedly differ from their 2011 counterparts.

tions nor removing all evaluations in 2012 by the 1,010 students who evaluated in both years affect our results qualitatively.

[12] We discuss the differences between male and female teachers shown in the last two columns of Table 1 in Section 6.

Table 2 compares the teachers in our sample with the 81 teachers who drop out of the sample after the first wave of student evaluations.[13] Attrition is balanced between the treatment and control group: 38 teachers (24.7%) drop out of the treatment group and 43 teachers (25.4%) drop out of the control group. Teachers who drop out of the sample receive lower student evaluations in the first wave as compared to teachers who remain in the sample. The difference is 0.11 points and statistically insignificant. The average self-assessment score is significantly lower among teachers who drop out as compared to the teachers in our sample. Furthermore, teachers who leave the sample are significantly older and have longer tenure, suggesting that retirement is partially responsible for attrition. The final two columns in Table 2 split the group of teachers who drop out by their assignment to the treatment group and the control group. Teachers who were assigned to the treatment group receive slightly worse student evaluation scores, evaluate themselves higher, and have longer tenure as compared to teachers assigned to the control group. However, none of these differences is statistically significant.[14]

Not all teachers in our sample completed the self-assessment questionnaire. Among the 242 teachers in our analysis, 166 teachers performed the self-assessment in the first year and 132 teachers did so in both years. Table 3 compares the teachers who completed the self-assessment survey twice with the teachers who did so only once or never. Most importantly, there is no significant difference between the treatment and control group in the number of times a teacher completes the self-evaluation. Furthermore, we find no difference in first-wave self-evaluation scores between teachers who did and did not complete the second self-evaluation. We do find that teachers who completed none of the self-evaluations receive significantly lower student evaluation scores in the first wave. On observables, males are relatively likely to refrain from completing the first self-evaluation.

4 Empirical strategy

We estimate the effect of receiving feedback using OLS with time- and teacher-fixed effects. The dependent variable, denoted by y_{it}, is the average student evaluation score of teacher i at time $t \in \{1, 2\}$. This is given by the average score on the 19 items on the evaluation questionnaire (see the Appendix) averaged over all students who evaluate the teacher in a given

[13] A large fraction of these 81 teachers left the school, in part as a result of a severance pay package offered to employees after the merger.

[14] We further examine the issue of selective attrition in Section 5.

year.[15] The main variable of interest is T_{it}, which is a dummy variable that equals one in the second year when teacher i is part of the treatment group and zero otherwise. Furthermore, we include teacher-fixed effects, denoted by θ_i, and time-fixed effects, by including dummy variable E_t that takes value 1 in the second year of our experiment and is zero otherwise. The regression equation reads:

$$y_{it} = \gamma T_{it} + \theta_i + \mu E_t + \varepsilon_{it}. \tag{1}$$

The estimated average treatment effect of receiving feedback is given by γ. Equation (1) is specified at the teacher level. We also estimate the average treatment effect at the student level. In all our estimations, we cluster standard errors at the teacher level.

Next, we investigate how the effect of receiving feedback depends on the content of the feedback, in two different ways. First, we include the interaction between the treatment dummy and the variable $\triangle self_i$, which denotes the difference between teacher i's average self-assessment score in the first year and teacher i's average student evaluation score in the first year. We analyze this interaction effect by estimating:

$$y_{it} = \gamma T_{it} + \varphi \left(T_{it} \times \triangle self_i\right) + \psi \left(E_t \times \triangle self_i\right) + \theta_i + \mu E_t + \varepsilon_{it}. \tag{2}$$

Note that we also interact $\triangle self_i$ with dummy variable E_t. This interaction accounts for correlations between second-year evaluation scores and $\triangle self_i$ that are independent of whether the teacher received her first-year evaluation scores, for instance due to reversion to the mean.

The relation between the content of feedback and subsequent performance may be non-linear. We perform a linear spline regression, allowing for different relations between the effect of feedback and $\Delta self_i$ for positive and negative values of $\Delta self_i$. Hence, we estimate:

$$\begin{aligned} y_{it} = {} & \gamma T_{it} + \varphi_p \left(T_{it} \times \triangle self_i^{pos}\right) + \varphi_n \left(T_{it} \times \triangle self_i^{neg}\right) \\ & + \psi_p \left(E_t \times \triangle self_i^{pos}\right) + \psi_n \left(E_t \times \triangle self_i^{neg}\right) + \theta_i + \mu E_t + \varepsilon_{it}, \end{aligned} \tag{3}$$

where $\triangle self_i^{pos} = \Delta self_i$ if $\Delta self_i > 0$ and $\triangle self_i^{pos} = 0$ if $\Delta self_i \leq 0$. Variable $\triangle self_i^{neg}$ correspondingly captures the negative values of $\Delta self_i$.[16]

Second, in a similar way we include the interaction between the treatment dummy and the variable $\triangle team_i$, which gives the difference between teacher

[15] Using the average score excluding statements 14 and 15 (as used to stratify assignment to treatment) does not affect our results in any important way.

[16] At $\Delta self_i = 0$, the teacher's and students' average assessment is identical, which makes it a natural level for the kink in the spline regression. None of our results is affected qualitatively when we impose that the kink is at any position in $[-0.5, 0.5]$.

i's average student evaluation score in the first year and the average of the first-year evaluations of all teachers in her team. Hence, $\triangle team_i$ denotes the extent to which teacher i performs better or worse than her colleagues, on average, as measured by the student evaluation scores.

Lastly, we estimate equations (1) to (3) using teachers' second-year average self-assessment scores and job satisfaction as dependent variables.

Our results are nearly identical if we standardize each questionnaire item before averaging (separately for student evaluations and teachers' self-assessment, and by year) and use the average standardized score as dependent variable. The reason is that if we standardize the variables of the 19 different items before averaging over them, we obtain variables that are highly correlated with the (non-standardized) variables we use in our analysis. For the average student evaluation score, this correlation is 0.98 in 2011 and 0.97 in 2012. For teachers' self-assessment, the correlation is 0.95 in 2011 and 0.93 in 2012. This carries over to the key interaction variables used in the analysis. If we take the difference between the average standardized self-assessment score and the average standardized student evaluation score, this variable has a correlation of 0.99 with the corresponding variable we use in our analysis, $\Delta self_i$. Similarly, we also find a correlation of 0.99 between our variable $\triangle team_i$ and the difference between a teacher's average standardized student evaluation score and the average standardized student score within the teacher's team. The advantage of using $\Delta self_i$ and $\triangle team_i$ is that these variables, in contrast to the standardized variables, naturally allow to distinguish between teachers whose evaluation scores are above rather than below their self-assessment score and average team scores.

5 Results

The estimates of the average treatment effect of receiving feedback on subsequent student evaluation scores are given in Table 4. The first column gives the results of estimating (1). The estimated average treatment effect on the average student evaluation score is 0.043, which is both economically and statistically insignificant.[17] This effect is quite precisely estimated, with a standard error equal to 0.054 and a 95 percent confidence interval that runs

[17] Given randomized assignment to treatment and control, it is possible to estimate the effect of feedback using only the data from the second wave of evaluations. Doing so yields a slightly higher but still statistically insignificant average treatment effect estimated at the teacher level: a coefficient of 0.085 with a standard error of 0.062. Including teacher-fixed effects has the advantage of increasing power.

from −0.063 to 0.149.[18] This result is in contrast to most previous studies on the provision of feedback as discussed in the Introduction, which usually find a positive effect of feedback on performance. A possible explanation for this difference is that previous studies typically focus on the effect of feedback in the short run, whereas we study the effect of feedback over the period of a full year. This interpretation is consistent with Azmat et al. (2019) who find that students respond to relative performance information in the short run, but not in the long run (where the long run in their paper is a full year, as in ours).

Ceiling effects may provide another possible explanation: given that student evaluations are quite high to start with, it may be (nearly) impossible to increase them. If so, we would expect to find larger average treatment effects on individual questionnaire items where teachers score relatively low in the first wave, such as items 3, 12, and 19 (see Table A.1 in the Appendix). Figure A.1 in the Appendix depicts the estimated average treatment effect on each of the 19 items of the questionnaire separately. Estimated effects range from 0.00 to 0.11, and is marginally significant (at the 0.06 level) only for item 5 ("The teacher is able to explain the connection to the real world."). The estimates do not show larger effects for items with low evaluation scores in the first year. Even on item 12 ("The teacher checks whether I did my assignments or homework."), which has the second-lowest evaluation score in 2011 and is arguably relatively easy to improve upon, we find no significantly larger response to feedback. Hence, we consider it unlikely that ceiling effects drive the lack of response on average.

Furthermore, Figure A.1 does not shows a clear pattern in terms of the 4 underlying categories (didactical skills (items 1 - 6), pedagogical skills (items 7 - 11), organizational aspects (items 12 - 15), and interpersonal skills (items (16 - 19)). Estimating the average treatment effect per category, by taking the average student evaluation score over the items per category as dependent variable, yields outcomes that are very close to the results in column 1 of Table 4.

The second column of Table 4 shows the average treatment effect estimated at the student level. Here, the dependent variable is the average evaluation score of a teacher by individual students. Again, the estimated average treatment effect is small and statistically insignificant. The differ-

[18] This non-significant result is not due to a lack of power. We can detect an effect of 0.076 (16 percent of a standard deviation) with 0.80 power. This is estimated as follows: we perform a regression of average student evaluation on year- and teacher-fixed effects. The residuals from this regression for observations in 2012 have a standard deviation of 0.20 in the treatment group and 0.21 in the control group. Using standard power calculations, this yields an minimally detectable effect size of 0.076.

ence between the two estimates indicates that the average treatment effect is slightly higher for teachers who are evaluated by relatively few students.[19] In the remainder of this paper, we only report the estimates at the teacher level; the estimated effects at the student level are qualitatively similar.

Next, we consider possible heterogeneity in treatment effects depending on the content of the feedback. First, we investigate whether the effect of feedback depends on the gap between teachers' self-assessment scores and the evaluation scores they receive from their students ($\Delta self_i$). Column 1 of Table 5 gives the results of estimating (2). If $\Delta self_i = 0$, the estimated treatment effect is very close to zero at 0.014. Hence, teachers who learn that their students' assessment is equal to their self-assessment hardly respond. The interaction effect between the treatment dummy and $\Delta self_i$ is positive but statistically insignificant. For teachers who learn that their students' evaluation score is one point lower than their self-assessment, the estimated treatment effect is $0.014 + 0.104 = 0.118$.

In column 2, we report the results of estimating (3), the specification that allows for non-linearity. To facilitate the interpretation of these results, Figure 4 depicts the estimated effects of receiving feedback as reported in column 2 of Table 5. We find that teachers whose own assessment corresponds to students' assessment do not respond to receiving feedback. As with the linear specification (2), the estimated treatment effect is positive for teachers who learn that their student evaluation score differs a lot from their self-assessment. This effect is significant at the 5-percent level for teachers whose self-assessment exceeds their average student evaluation scores by more than one point. However, the fraction of teachers in this interval is fairly small, about ten percent (as can be seen from light grey kernel density in Figure 4).[20]

Second, we examine whether the effect of feedback depends on the gap between a teacher's first-period student evaluation score and the average

[19] In the estimation at the teacher level, all teachers are weighted equally, independent of the number of students that evaluate them. In contrast, teachers who are evaluated by many students receive a higher weight in the estimation at the student level, relative to teachers who are evaluated by few students. Estimating the average treatment effect at the teacher level while weighing teachers by the number of students evaluating them in either the first or second wave gives results close to those reported in column 2 of Table 4.

[20] We also examined whether treatment effects differ by first-period student evaluation score. To do so, we ran a regression similar to (2), but with first-period student evaluation score instead of $\triangle self_i$. We find that the treatment effect is very close to zero and negatively but not significantly related to first-period student evaluation score. Including both first-period student evaluation score and $\triangle self_i$ in one single regression gives rise to problems of multicollinearity. The correlation between first-period student evaluation score and $\triangle self_i$ is -0.71.

score in his team. The third column of Table 5 gives the results of estimating (2) with $\triangle team_i$ instead of $\triangle self_i$. We find that the estimated interaction effect is negative and statistically insignificant. The estimated treatment effect for teachers who learn that they perform as well as their team (on average) is 0.062. For teachers who learn that their student evaluation score is one point above the average of their colleagues, this effect is reduced by 0.090 points. In column 4, we report the results of estimating (3), allowing for different relations between the effect of feedback and $\Delta team_i$ for positive and negative values of $\Delta team_i$. As illustrated in Figure 5, the estimated treatment effect is positive for teachers who learn they perform worse than their teams' average, but only significant for teachers who learn that they score slightly worse than their colleagues (up to 0.5 points below their teams' average).[21]

The bottom half of Table 5 shows that among teachers in the control group a significant positive (negative) correlation between the average student evaluation score and $\triangle self_i$ ($\Delta team_i$) in the previous year exists. One explanation is that teachers sense that students are relatively dissatisfied, and try to improve this in the subsequent year. Moreover, regression to the mean likely explains a large part of these correlations: if average student evaluation scores contain a random component, teachers with relatively high (low) scores are likely to receive lower (higher) scores in the subsequent year. These correlations show the importance of introducing experimental variation, as the estimated effects of any school-wide policy change would be confounded by these time effects.

Figures A.2 and A.3 depict the results of estimating equation (3) for each questionnaire item separately. For this, we have created item-specific versions of variables $\triangle self_i$ and $\Delta team_i$. The kernel densities of these item-specific variables are given by the shaded area in each plot in Figures A.2 and A.3. Figure A.2 shows that the key effect highlighted in Figure 4 (a positive treatment effect for teachers who learn that their average student evaluation score is substantially lower than their average self-assessment score) is driven by questionnaire items 2 to 7, 14, and 19. Many of these belong to the category didactical skills (items 1 to 6). The organizational aspects contribute little to the pattern in Figure 4, except for answering emails on time (item 14). Similarly, Figure A.3 shows that the pattern depicted in Figure 5 is mostly driven by items 1 to 5, 11, 12, and 19. Hence, teachers who learn that they score somewhat lower than their colleagues on didactical skills items 1 to 5

[21] Other specifications present a similar picture, including quadratic splines and estimating treatment effects for subsets of values of $\Delta self_i$ and $\Delta team_i$. Results are available upon request.

improve modestly. The same holds for the clarity of expectations and checking homework (items 11 and 12) and in setting a good example (item 19). For organizational aspects, we again find no clear effects of our treatment.

As discussed before, 81 teachers who were evaluated in 2011 and assigned to either the treatment group or control group were not evaluated in 2012 and, hence, are not included in the analysis. In Section 3, we showed that attrition is unrelated to being assigned to the treatment group and also unrelated to student evaluation scores in the first wave (see Table 2). However, if attrition is related to the content of the feedback received, the teachers who drop out of the treatment group may differ from the teachers who drop out of the control group, which could bias our results. To examine whether attrition is related to the content of the feedback received, we perform regressions on the set of teachers with student evaluation scores in 2011, with as dependent variable a dummy that takes value 1 if a teacher drops out. As reported in Table A.3 in the Appendix, the estimations show that neither receiving feedback nor the content of this feedback significantly affects the probability of dropping out, with one exception. Column 2 shows that teachers who learn that their students evaluate them considerably better than their self-assessment are more likely to leave. As this result is based on only 5 leaving teachers for whom $\triangle self_i < 0$, we are confident that this does not affect our main results.[22]

We have seen that on average, teachers' self-assessment is much more favorable than the evaluations by their students. Hence, feedback on student evaluation score may help teachers in making a more realistic assessment of their own performance. As teachers were asked to complete the self-assessment in both waves, we can examine whether teachers use the feedback to update the self-assessment of their performance. Table 6 reports the effects of receiving feedback on teachers' self-assessment. The estimation reported in the first column only includes a treatment dummy, a year dummy, and teacher-fixed effects. We find that, on average, teachers who have received feedback evaluate themselves worse in the second wave compared to teachers who have not received feedback. The average treatment effect is -0.067, but statistically insignificant. The estimation reported in the second column adds the interaction between the treatment dummy and the difference between teachers' first-period self-assessment score and their students' first-period evaluation scores ($\triangle self_i$). As expected, the interaction effect is negative, but it is statistically insignificant. In column 3, we allow the inter-

[22] These results are robust to not including individual controls. Since we miss data on one or more individual characteristics for 41 teachers, the sample size then increases to 323.

action effect to differ for positive and negative values of $\triangle self_i$. As depicted in Figure 6, we find no significant effect of the treatment for teachers who learn that their student evaluation scores are higher than their self-assessed scores. In contrast, teachers who learn that their students' evaluation is less positive than their self-evaluation do assess themselves significantly less positive in the second wave, compared to similar teachers who do not receive feedback. Still, the magnitude of this adjustment is rather limited: about one tenth of a point for each full point the average student evaluation score exceeds the self-assessment.

Lastly, we examine whether receiving feedback affects teachers' job satisfaction. Teachers may be positively or negatively surprised about their average evaluation score, leading to feelings of pride or resentment. Similarly, learning that one's performance is better or worse than the performance of direct colleagues may affect job satisfaction as a result of status concerns or conformity preferences. The estimation reported in the first column of Table 7 includes only the treatment dummy.[23] We find that on average, receiving feedback has no effect on job satisfaction. The estimated effect is -0.068 (on a 5-point scale) and statistically insignificant. The estimation in the second column adds an interaction between the treatment dummy and $\triangle self_i$. Surprisingly, the estimated interaction effect is positive, but insignificant. Column 3 estimates the relation separately for positive and negative values of $\triangle self_i$. The results of this estimation are depicted in Figure 7. The effect of receiving feedback is very close to zero (except for teachers learning that student evaluation scores are much higher than their self-assessed score), but nowhere statistically significant.

In column 4 of Table 7, we interact the treatment dummy with the difference between a teacher's first-period average student evaluation score and her team's average student evaluation score ($\Delta team_i$). The estimated interaction effect is negative and insignificant. This also holds when we estimate this relation separately for positive and negative values of $\Delta team_i$ in column 5. Figure 8 depicts the results of the latter estimation. The estimated effect of receiving feedback on job satisfaction is close to zero for teachers whose evaluation scores are above their teams' average. For teachers who learn they perform worse than their direct colleagues, the estimated effect is positive, but not statistically significant. Hence, we find no effect of performance feedback on job satisfaction.

[23] Recall that we only have data on job satisfaction at one point in time. Hence, we cannot include teacher-fixed effects and a time-fixed effect. Instead, we include a set of controls (teachers' gender, age, tenure, and work time).

6 Gender differences in response to feedback

In this section, we analyze whether male and female teachers respond differently to receiving feedback. This analysis of gender differences was not planned in advance, but initiated following questions received from conference and seminar audiences when presenting the results shown in the previous section. Hence, this is an ex post, exploratory analysis, and the results should be interpreted as such.[24] In our final sample, we have 123 men and 112 women; for 7 teachers we have no information about gender. Table 1 compares characteristics of male and female teachers. On average, female teachers are three years younger than male teachers, have three years less tenure, and have considerably smaller contract sizes. Among the teachers who performed the first self-evaluation, male and female teachers rate themselves equally high. In the first survey among students, female teachers receive somewhat higher average evaluation scores than male teachers (4.17 versus 4.06), although the difference is not statistically significant. In a regression, controlling for age, tenure, and work time, the coefficient on the female dummy is 0.12, with a p-value of 0.066 (regression output not reported for brevity). On the sub-items of the student evaluation questionnaire, we do find that female teachers score significantly higher on items 12 to 15, which capture administrative organization.[25] None of these findings is affected when including the teachers who were only evaluated in the first survey.

To determine whether men and women respond differently to feedback, we estimate equations (1) and (3) separately for male and female teachers. The regression results can be found in Table 8, and are depicted in Figures 9 to 11. The shaded areas in Figures 9 and 10 give the kernel densities of $\Delta self_i$ and $\Delta team_i$, respectively, separated by gender. For both variables, we find no statistically significant gender differences, neither in the average value nor in the distributions (using Kolmogorov-Smirnov tests). Hence, any gender differences are not driven by differences in the content of information across genders. Note that in these estimations, we do not control for the interactions between receiving feedback and other characteristics. If we do

[24] Our data do not allow us to examine gender bias in student evaluations. Recently, Boring (2017) and Mengel et al. (2019) find that female teachers receive lower student evaluation scores than male teachers, despite being equally effective in terms of student performance on standardized tests. In our data, student evaluation scores do not differ significantly between male and female teachers (see Table 1), but this obviously does not rule out gender bias.

[25] In the self-assessment, female teachers do rate themselves significantly higher on item 12, but not on the other items.

control for the interaction between the content of feedback and other observable characteristics (age, tenure, and work time), we find similar results. Of course, it is possible that the gender differences are (partially) driven by non-observed factors, leading to omitted variable bias.

Columns 1 and 2 of Table 8 give the results of estimating the average treatment effect of receiving feedback on subsequent student evaluation scores. Female teachers respond more strongly to receiving feedback than male teachers, although the difference is not statistically significant. Columns 3 and 4 give the results of interacting the treatment with $\Delta self_i^{pos}$ and $\Delta self_i^{neg}$. As depicted in Figure 9, our finding that teachers do respond to receiving 'bad news' can be entirely attributed to female teachers. Male teachers do not respond to learning that their student evaluation scores are lower than their self-assessment score, whereas female teachers' subsequent student evaluation scores increase significantly.[26] We obtain a similar result when replacing $\Delta self_i$ with $\Delta team_i$ in columns 5 and 6, depicted in Figure 10. Women do respond to receiving a student evaluation score below their teams' average. Men's response, in contrast, is entirely independent of how their score differs from the score of their direct colleagues. Columns 7 and 8 and Figure 11 show that these findings carry over to the effect of receiving feedback on self-assessment. Men's self-assessment is not affected at all when receiving student evaluation scores below their self-evaluation scores. Women do show a downward adjustment in their self-evaluation after receiving relatively low student evaluation scores. Finally, we do not find any gender differences in the relation between job satisfaction and receiving feedback (regression results not reported for brevity). Hence, in short, whereas male teachers by and large seem to ignore the feedback provided, female teachers do respond depending on the content of feedback.

To provide more insight into why women appear to respond more strongly than men, Figures A.4 to A.6 present the same analyses by questionnaire item. Figure A.4 depicts item-specific average treatment effects by gender. None of the estimated effects differs significantly from zero. The estimated effect is consistently higher for females than for males, but the difference is statistically significant (at the 10 percent level) only for items 13, 15, and 16. Figure A.5 gives the item-specific interaction effects with $\Delta self_i^{pos}$ and $\Delta self_i^{neg}$ for females and males separately. Males are consistently unresponsive to feedback, particularly to student feedback that is worse than their self-assessment. In contrast, females' response to 'bad news' is driven by

[26] Interestingly, both male and female teachers seem to respond positively to 'very good news', i.e. learning that student evaluation scores are much higher than their self-assessment scores. As is clear from Figure 9, this applies to only a very small share of the sample.

items 1 to 9, 14 to 17, and 19. Hence, the gender difference in response to negative feedback is by and large consistent across all four teaching aspects. Similarly, Figure A.6 shows the item-specific interaction effects with $\Delta team_i^{pos}$ and $\Delta team_i^{neg}$ by gender. Again, we find that males are consistently unresponsive to learning that they receive worse evaluations than their colleagues. Females' responses vary across items, with particularly strong responses for items 1, 5, 10, 11, and 15. The response is less strong for items 16 to 19, indicating that females respond less to learning that they receive lower scores than their colleagues on interpersonal skills, compared to learning this regarding other teaching aspects.

7 Conclusion

This paper has studied whether and when teachers respond to receiving students' feedback by conducting a large-scale field experiment in vocational education. We find that on average, teachers to not respond to receiving students' feedback. This result contrasts with recent studies about performance feedback, which tend to find positive effects in the short run (typically within a semester). One reason for the difference in results might be that effects of feedback are short-lived. A possible remedy for this problem might be to provide feedback more frequently. It would be interesting to examine in a future field experiment how teachers respond to more frequent feedback, and to learn about the dynamics of this response. There could also be other reasons for the lack of an average treatment effect in our study, e.g. because teacher-student matches are changing or because teachers in our context face no incentive to improve their student evaluation scores. A future meta-analysis of studies on student feedback could shed light on how important these contextual features are for the effectiveness of student feedback systems.

Additionally, we examined whether the response to feedback depends on the content of feedback. We found that teachers who learn that their students' assessment is much less favorable than their own assessment improve student evaluation scores after receiving feedback. These teachers also moderate their self-assessment, albeit to a limited extent. Teachers who learn that they are evaluated worse as compared to the average score in their team improve, albeit to a limited extent. We found no evidence that teachers' job satisfaction is affected by (the content of) feedback. These content-dependent responses to receiving feedback appear to be entirely driven by female teachers, while male teachers hardly respond to any feedback. As the latter finding is based on explorative ex post analysis, further research is needed to validate

this result.

References

[1] Auriol, Emmanuelle, and Régis Renault (2008), Status and Incentives, *RAND Journal of Economics*, 39(1): 305-326.

[2] Azmat, Ghazala, Manuel Bagues, Antonio Cabrales, and Nagore Iriberri (2019), What You Know Can't Hurt You? A Field Experiment on Relative Feedback Performance, *Management Science,* 65(8): 3714-3736.

[3] Azmat, Ghazala, and Nagore Iriberri (2010), The Importance of Relative Performance Feedback Information: Evidence from a Natural Experiment Using High School Students, *Journal of Public Economics*, 94(7): 435-452.

[4] Azmat, Ghazala, and Nagore Iriberri (2016), The Provision of Relative Performance Feedback Information: An Analysis of Performance and Happiness, *Journal of Economics & Management Strategy*, 25(1): 77-110.

[5] Bandiera, Oriana, Iwan Barankay, and Imran Rasul (2013), Team Incentives: Evidence from a Firm Level Experiment, *Journal of the European Economic Association*, 11(5): 1079-1114.

[6] Barankay, Iwan (2012), Rank Incentives: Evidence from a Randomized Workplace Experiment, Mimeo.

[7] 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.

[8] Benabou, Roland, and Jean Tirole (2006), Incentives and Prosocial Behavior, *American Economic Review*, 96(5): 1652-1678.

[9] Bernheim, Douglas B. (1994), A Theory of Conformity, *Journal of Political Economy*, 102(5): 841-877.

[10] Besley, Timothy, and Maitreesh Ghatak (2005), Competition and Incentives with Motivated Agents, *American Economic Review*, 95(3): 616-636.

[11] Besley, Timothy, and Maitreesh Ghatak (2008), Status Incentives, *American Economic Review*, 98(2): 206-211.

[12] Blanes i Vidal, Jordi, and Mareike Nossol (2011), Tournaments Without Prizes: Evidence from Personnel Records, *Management Science*, 57(10): 1721-1736.

[13] Boring, Anne (2017), Gender Biases in Student Evaluations of Teaching, *Journal of Public Economics* 145: 27-41.

[14] Braga, Michela, Marco Paccagnella, and Michele Pellizzari (2014), Evaluating Students' Evaluations of Professors, *Economics of Education Review,* 41: 71-88.

[15] Briole, Simon, and Eric Maurin (2019), Does Evaluating Teachers Make a Difference?, IZA Discussion Paper No. 12307.

[16] Burgess, Simon, Shenila Rawal, and Eric S. Taylor (2019), Teacher Peer Observation and Student Test Scores: Evidence from a Field Experiment in English Secondary Schools, Harvard Graduate School of Education Working Paper.

[17] Buser, Thomas, Leonie Gerhards, and Joël van der Weele (2018), Responsiveness to Feedback as a Personal Trait, *Journal of Risk and Uncertainty,* 56(2): 165-192.

[18] Carrell, Scott E., and James E. West (2010), Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors, *Journal of Political Economy*, 118(3): 409-432.

[19] Centra, John A. (1973), Effectiveness of Student Feedback in Modifying College Instruction, *Journal of Educational Psychology*, 65(3): 395-401.

[20] Cohen, Peter A. (1980), Effectiveness of Student-Rating Feedback for Improving College Instruction: A Meta Analysis of Findings, *Research in Higher Education*, 13(4): 321-341.

[21] Delfgaauw, Josse, and Robert Dur (2008), Incentives and Workers' Motivation in the Public Sector, *Economic Journal*, 118: 171-191.

[22] Delfgaauw, Josse, Robert Dur, Joeri Sol, and Willem Verbeke (2013), Tournament Incentives in the Field: Gender Differences in the Workplace, *Journal of Labor Economics*, 31(2): 305-326.

[23] Gerhards, Leonie, and Neele Siemer (2016), The Impact of Private and Public Feedback on Worker Performance: Evidence from the Lab, *Economic Inquiry* 54(2): 1188-1201.

[24] Johnson, Maria, and Vicki S. Helgeson (2002), Sex Differences in Response to Evaluative Feedback: A Field Study, *Psychology of Women Quarterly* 26(3): 242-251.

[25] Kuhnen, Camelia M., and Agnieszka Tymula (2012), Feedback, Self-Esteem and Performance in Organizations, *Management Science*, 58(1): 94-113.

[26] List, John A., Sally Sadoff, and Mathis Wagner (2011), So You Want To Run an Experiment, Now What? Some Simple Rules of Thumb For Optimal Experimental Design, *Experimental Economics,* 14(4): 439-457.

[27] Marsh, Herbert W. (2007), Students' Evaluations of University Teaching: Dimensionality, Reliability, Validity, Potential Biases and Usefulness. In R.P Perry and J.C. Smart (Eds.), The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective, pp. 319-383. New York: Springer.

[28] Mengel, Friederike, Jan Sauermann, and Ulf Zölitz (2019), Gender Bias in Teaching Evaluations, *Journal of the European Economic Association*, 17(2): 535-566.

[29] Möbius, Markus M., Muriel Niederle, Paul Niehaus, and Tanya S. Rosenblat (2007), Gender Differences in Incorporating Performance Feedback, Mimeo, Harvard University.

[30] Moldovanu, Benny, Aner Sela, and Xianwen Shi (2007), Contests for Status, *Journal of Political Economy*, 115(2): 338-363.

[31] Roberts, Tomi-Ann, and Susan Nolen-Hoeksema (1994), Gender Comparisons in Responsiveness to Others' Evaluations in Achievement Settings, *Psychology of Women Quarterly* 18(2): 221–240.

[32] Ryan, James J., James A. Anderson, and Allen B. Birchler (1980), Student Evaluation: The Faculty Responds, *Research in Higher Education,* 12(4): 317-333.

[33] Sliwka, Dirk (2007), Trust as a Signal of a Social Norm and the Hidden Costs of Incentive Schemes, *American Economic Review*, 97(3): 999-1012.

[34] Suurmond, Guido, Otto Swank, and Bauke Visser (2004), On the Bad Reputation of Reputational Concerns, *Journal of Public Economics*, 88(12): 2817-2838.

[35] Taylor, Eric S., and John H. Tyler (2012), The Effect of Evaluation on Teacher Performance, *American Economic Review,* 102(7): 3628-3651.

[36] Tran, Anh, and Richard Zeckhauser (2012), Rank as an Inherent Incentive: Evidence from a Field Experiment, *Journal of Public Economics*, 96(9): 645-650.

[37] Watts, Michael, and William E. Becker (1999), How Departments of Economics Evaluate Teaching, *American Economic Review*, 89(2): 344-349.

Tables

Table 1: Descriptive statistics of teachers

	Treatment group	Control group	All teachers	Male teachers	Female teachers
First wave evaluation by students					
Mean	4.15	4.10	4.12	4.06	4.17
Standard deviation	(0.46)	(0.49)	(0.48)	(0.47)	(0.49)
First wave number of evaluations by students					
Mean	32.27	33.40	32.86	33.19	32.78
Standard deviation	(12.65)	(14.97)	(13.89)	(15.09)	(12.45)
First wave self-evaluation[a]					
Mean	4.62	4.59	4.60	4.58	4.62
Standard deviation	(0.29)	(0.30)	(0.29)	(0.27)	(0.31)
Gender: % Female					
Mean	0.46	0.49	0.48		
Standard deviation	(0.50)	(0.50)	(0.50)		
Age: Years					
Mean	47.25	49.22	48.26	49.50	46.8^{++}
Standard deviation	(10.26)	(9.97)	(10.14)	(9.90)	(10.26)
Employment: % of fte					
Mean	0.76	0.81*	0.78	0.88	0.68^{+++}
Standard deviation	(0.23)	(0.20)	(0.21)	(0.16)	(0.21)
Tenure: Years					
Mean	14.10	16.42*	15.28	16.89	13.53^{++}
Standard deviation	(10.42)	(10.01)	(10.26)	(10.17)	(10.11)
Number of teachers	116	126	242	123	112

Notes: [a] The self-evaluation was completed by 166 teachers in our sample, 82 in the treatment group and 84 in the control group, 79 males, and 85 females. *, **, and *** indicate a statistically significant difference between the treatment group and control group at the .10, .05, and .01 level, respectively. $^{+}$, $^{++}$, and $^{+++}$ indicate a statistically significant difference between male teachers and female teachers at the .10, .05, and .01 level, respectively.

Table 2: Attrition

	Sample	Total Attrition	Attrition Treatment	Attrition Control
First wave evaluation by students				
Mean	4.12	4.02	3.98	4.06
Standard deviation	(0.48)	(0.56)	(0.58)	(0.56)
First wave number of evaluations by students				
Mean	32.27	28.15**	27.34	28.86
Standard deviation	(12.65)	(15.58)	(14.71)	(16.46)
First wave self-evaluation[a]				
Mean	4.60	4.41***	4.51	4.36
Standard deviation	(0.29)	(0.68)	(0.35)	(0.81)
Gender: % Female				
Mean	0.48	0.47	0.48	0.46
Standard deviation	(0.50)	(0.50)	(0.51)	(0.51)
Age: Years				
Mean	48.26	50.95*	50.96	50.95
Standard deviation	(10.14)	(9.75)	(8.65)	(10.50)
Employment: % of fte				
Mean	0.78	0.76	0.74	0.77
Standard deviation	(0.21)	(0.23)	(0.20)	(0.25)
Tenure: Years				
Mean	15.28	18.15*	16.58	19.17
Standard deviation	(10.26)	(10.16)	(9.90)	(10.37)
Number of teachers	242	81	38	43

Notes: [a] The self-evaluation was completed by 166 teachers in our sample and by 46 teachers who dropped out, of whom 29 had been assigned to the treatment group and 17 to the control group. *, **, and *** indicate a statistically significant difference between the sample group and attrition group at the .10, .05, and .01 level, respectively. Within the group of teachers who drop out, none of the differences between teachers assigned to the treatment group and teachers assigned to the control group are statistically signficant.

Table 3: Descriptive statistics by self-evaluation

	Both self-evaluations completed	Only the first self-evaluation completed	No self-evaluation completed
Assigned to treatment group			
Mean	0.49	0.50	0.45
Standard deviation	(0.50)	(0.51)	(0.50)
First wave evaluation by students			
Mean	4.21	4.05*	4.00^{+++}
Standard deviation	(0.41)	(0.49)	(0.55)
First wave number of evaluations by students			
Mean	32.20	33.41	33.74
Standard deviation	(13.05)	(12.49)	(15.88)
First wave self-evaluation			
Mean	4.60	4.61	
Standard deviation	(0.30)	(0.27)	
Gender: % Female			
Mean	0.50	0.61	$0.37^{+,@@}$
Standard deviation	(0.50)	(0.50)	(0.49)
Age: Years			
Mean	48.23	46.88	48.94
Standard deviation	(9.91)	(11.21)	(10.12)
Employment: % of fte			
Mean	0.82	0.69***	0.77
Standard deviation	(0.19)	(0.24)	(0.22)
Tenure: Years			
Mean	15.42	15.53	14.89
Standard deviation	(9.97)	(11.35)	(10.41)
Number of teachers	132	34	76

Notes: *, **, and *** indicate a statistically significant difference between column (1) and (2) at the .10, .05, and .01 level, respectively.
$^{+}$, $^{++}$, and $^{+++}$ indicate a statistically significant difference between column (1) and (3) at the .10, .05, and .01 level, respectively.
$^{@}$, $^{@@}$, and $^{@@@}$ indicate a statistically significant difference between column (2) and (3) at the .10, .05, and .01 level, respectively.

Table 4: Effect of feedback on teachers' performance

Dependent variable: average student evaluation	(1) Teacher level	(2) Student level
Treatment	0.043	0.021
	(0.054)	(0.046)
Year dummy	Yes	Yes
Teacher-fixed effects	Yes	Yes
Observations	484	15194
Teachers	242	242
Overall R^2	0.006	0.002

Notes: Standard errors clustered at the teacher level between parentheses. *, **, and *** indicate significance based on a two-sided test at the .10, .05, and .01 level, respectively.

Table 5: Heterogenous treatment effects of feedback on performance

Dependent variable: average student evaluation				
	(1)	(2)	(3)	(4)
Treatment	0.014	-0.067	0.062	0.105
	(0.066)	(0.091)	(0.048)	(0.076)
Δself × treatment	0.104			
	(0.110)			
Δself+ × treatment		0.207		
		(0.140)		
Δself− × treatment		-0.275		
		(0.231)		
Δteam × treatment			-0.090	
			(0.101)	
Δteam+ × treatment				-0.227
				(0.225)
Δteam− × treatment				0.001
				(0.167)
Δself × second period	0.216**			
	(0.087)			
Δself+ × second period		0.294**		
		(0.115)		
Δself− × second period		-0.105		
		(0.168)		
Δteam × second period			-0.369***	
			(0.075)	
Δteam+ × second period				-0.314**
				(0.154)
Δteam− × second period				-0.407***
				(0.134)
Teacher fixed effects	Yes	Yes	Yes	Yes
Year dummy	Yes	Yes	Yes	Yes
Observations	332	332	484	484
Teachers	166	166	242	242
Overall R^2	.098	.086	.122	.119

Notes: Standard errors clustered at the teacher level between parentheses. *, **, and *** indicate significance based on a two-sided test at the .10, .05, and .01 level, respectively. Variable Δself is the difference between a teacher's first-period average self-assessment score and her first-period average student evaluation score. Variable Δteam is the difference between a teacher's first-period average student evaluation score and the average of all first-period average student evaluation scores of the teachers in her team.

Table 6: Effect of feedback on the teachers' self-evaluation

Dependent variable: average self-evaluation	(1)	(2)	(3)
Treatment	-0.067	-0.042	-0.040
	(0.046)	(0.059)	(0.076)
Δself $\times$ treatment		-0.108	
		(0.097)	
Δself+ $\times$ treatment			-0.104
			(0.122)
Δself$-$ $\times$ treatment			0.012
			(0.230)
Δself $\times$ second period		-0.091	
		(0.060)	
Δself+ $\times$ second period			-0.014
			(0.075)
Δself$-$ $\times$ second period			-0.418***
			(0.120)
Teacher fixed effects	Yes	Yes	Yes
Year dummy	Yes	Yes	Yes
Observations	264	264	264
Teachers	132	132	132
Overall R^2	.002	.031	.051

Notes: Standard errors clustered at the teacher level between parentheses. *, **, and *** indicate significance based on a two-sided test at the .10, .05, and .01 level, respectively. Variable Δself is the difference between a teacher's first-period average self-assessment score and her first-period average student evaluation score.

Table 7: Effect of feedback on teachers' job satisfaction

Dependent variable: Job satisfaction					
	(1)	(2)	(3)	(4)	(5)
Treatment	-0.068	-0.257	-0.263	-0.047	-0.189
	(0.133)	(0.214)	(0.262)	(0.134)	(0.233)
Δself		-0.407			
		(0.262)			
Δself+			-0.423		
			(0.278)		
Δself−			0.884		
			(1.789)		
Δself × treatment		0.292			
		(0.321)			
Δself+ × treatment			0.301		
			(0.375)		
Δself− × treatment			-0.638		
			(1.931)		
Δteam				0.274	
				(0.211)	
Δteam+					-0.172
					(0.446)
Δteam−					0.544
					(0.361)
Δteam × treatment				-0.419	
				(0.283)	
Δteam+ × treatment					-0.013
					(0.650)
Δteam− × treatment					-0.777
					(0.498)
Individual controls	Yes	Yes	Yes	Yes	Yes
Observations	162	130	130	162	162
R^2	.032	.052	.024	.047	.025

Notes: Standard errors between parentheses. *, **, and *** indicate significance based on a two sided test at the .10, .05, and .01 level, respectively. Individual controls are gender, age, tenure, and full-time equivalent. Variable Δself is the difference between a teacher's first-period average self-assessment score and her first-period average student evaluation score. Variable Δteam is the difference between a teacher's first-period average student evaluation score and the average of all first-period average student evaluation scores of the teachers in her team.

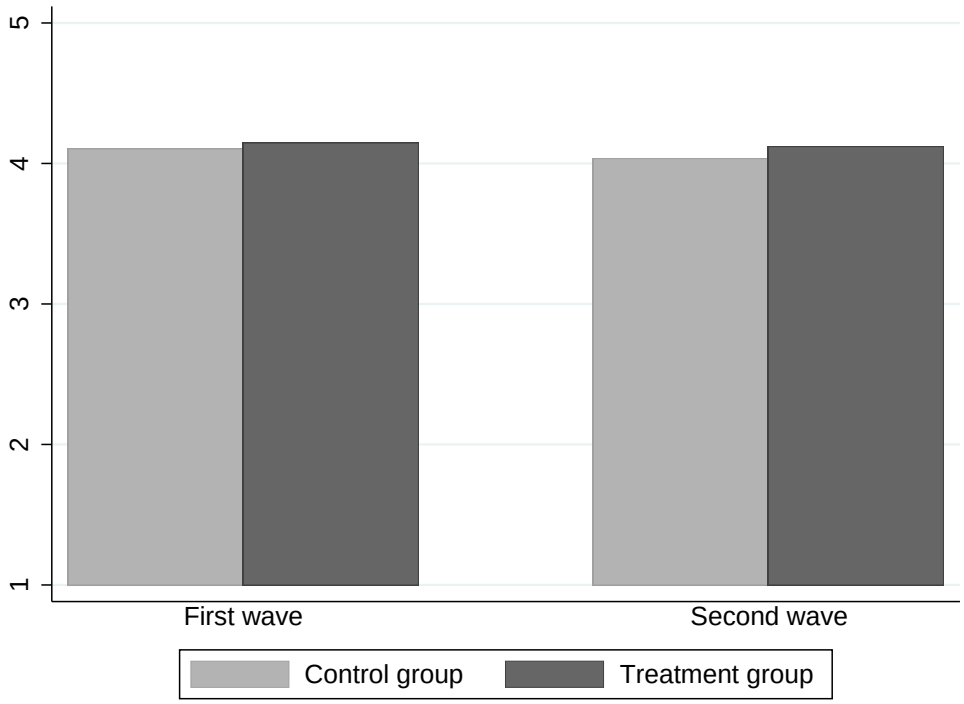
Table 8: Gender differences in the effects of feedback

Dependent variable:	avg student evaluation		avg student evaluation		avg student evaluation		self-evaluation	
	Male	Female	Male	Female	Male	Female	Male	Female
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	-0.043	0.099	-0.053	-0.116	-0.007	0.214*	-0.078	0.038
	(0.067)	(0.088)	(0.147)	(0.121)	(0.102)	(0.115)	(0.110)	(0.098)
Δself+ × treatment			0.035	0.403**			-0.008	-0.230
			(0.230)	(0.161)			(0.144)	(0.216)
Δself− × treatment			-0.707*	-0.448			0.341	0.187
			(0.420)	(0.349)			(0.440)	(0.333)
Δteam+ × treatment					0.116	-0.797*		
					(0.271)	(0.420)		
Δteam− × treatment					-0.086	-0.163		
					(0.243)	(0.227)		
Δself+ × second period			0.312*	0.194			-0.051	0.058
			(0.176)	(0.132)			(0.099)	(0.133)
Δself− × second period			0.561	-0.185			-0.862**	-0.340***
			(0.384)	(0.180)			(0.403)	(0.092)
Δteam+ × second period					-0.377*	-0.344		
					(0.195)	(0.260)		
Δteam− × second period					-0.454***	-0.097		
					(0.159)	(0.182)		
Teacher fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year dummy	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	246	224	158	172	246	224	136	132
Teachers	123	112	79	86	123	112	68	66
Overall R^2	.010	.004	.130	.037	.156	.047	0.035	0.016

Notes: Standard errors clustered at the teacher level between parentheses. *, **, and *** indicate significance based on a two-sided test at the .10, .05, and .01 level, respectively. Variable Δself is the difference between a teacher's first-period average self-assessment score and her first-period average student evaluation score. Variable Δteam is the difference between a teacher's first-period average student evaluation score and the average of all first-period average student evaluation scores of the teachers in her team.

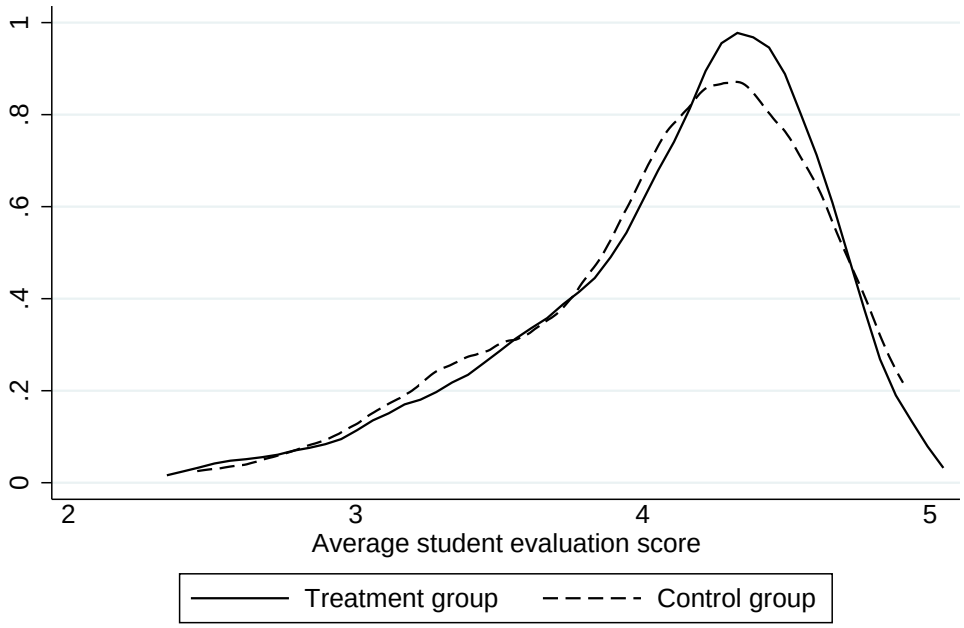
Figures

Figure 1: Average student evaluation scores by year



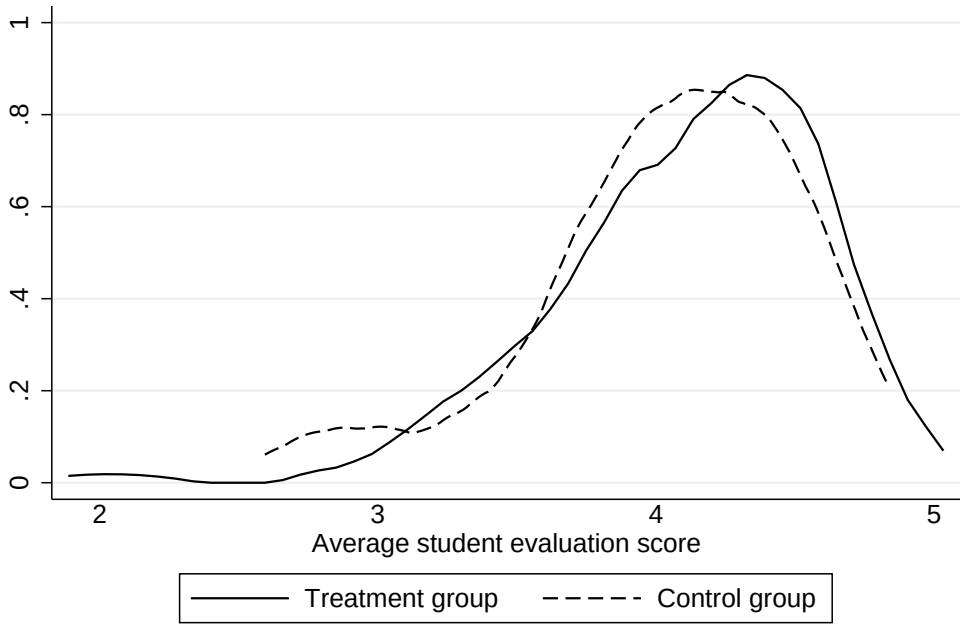
Notes: A student's evaluation of a teacher is defined as the average score on 19 statements on the teacher's performance (see the Appendix). The answer categories for each statement are [1] Disagree, [2] Disagree somewhat, [3] Disagree somewhat/Agree somewhat, [4] Agree somewhat, and [5] Agree.

Figure 2: Distribution of student evaluations in the first wave



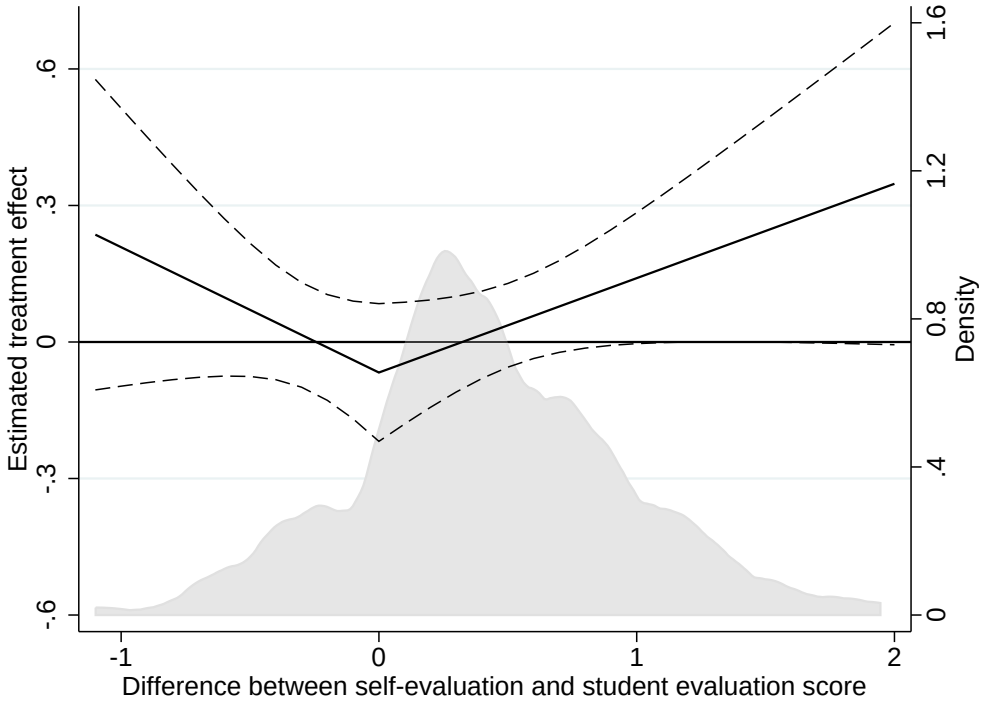
Notes: Distribution estimated using a kernel density function. A student's evaluation of a teacher is defined as the average score on 19 statements on the teacher's performance (see the Appendix). The answer categories for each statement are [1] Disagree, [2] Disagree somewhat, [3] Disagree somewhat/Agree somewhat, [4] Agree somewhat, and [5] Agree.

Figure 3: Distribution of student evaluations in the second wave



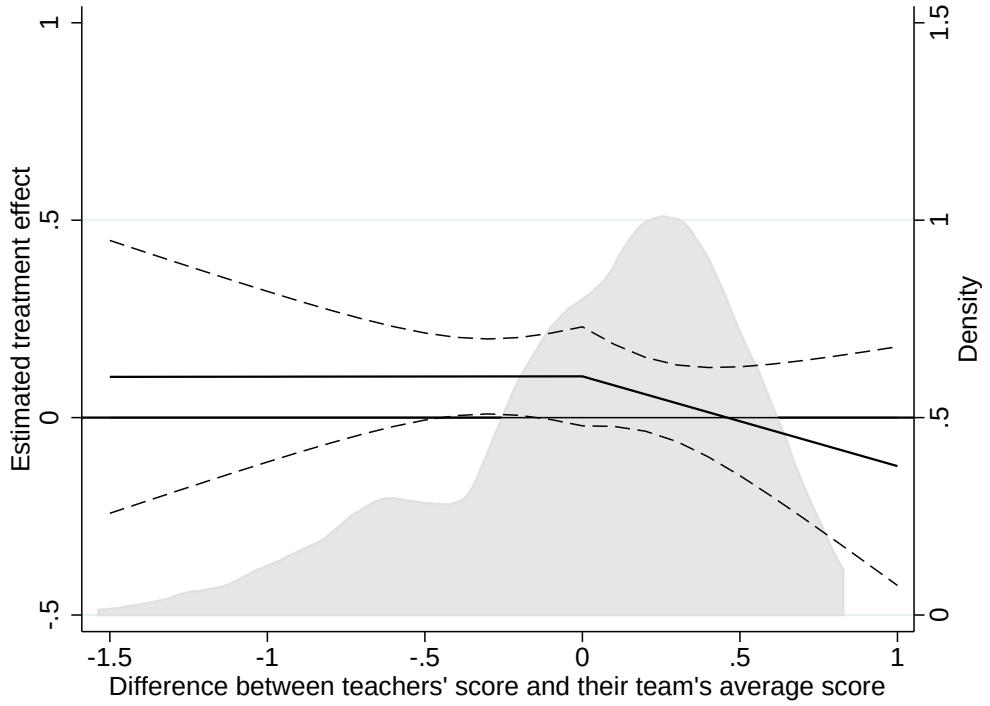
Notes: Distribution estimated using a kernel density function. A student's evaluation of a teacher is defined as the average score on 19 statements on the teacher's performance (see the Appendix). The answer categories for each statement are [1] Disagree, [2] Disagree somewhat, [3] Disagree somewhat/Agree somewhat, [4] Agree somewhat, and [5] Agree.

Figure 4: Estimated effect of feedback by the difference between a teacher's self-evaluation score and her student evaluation score



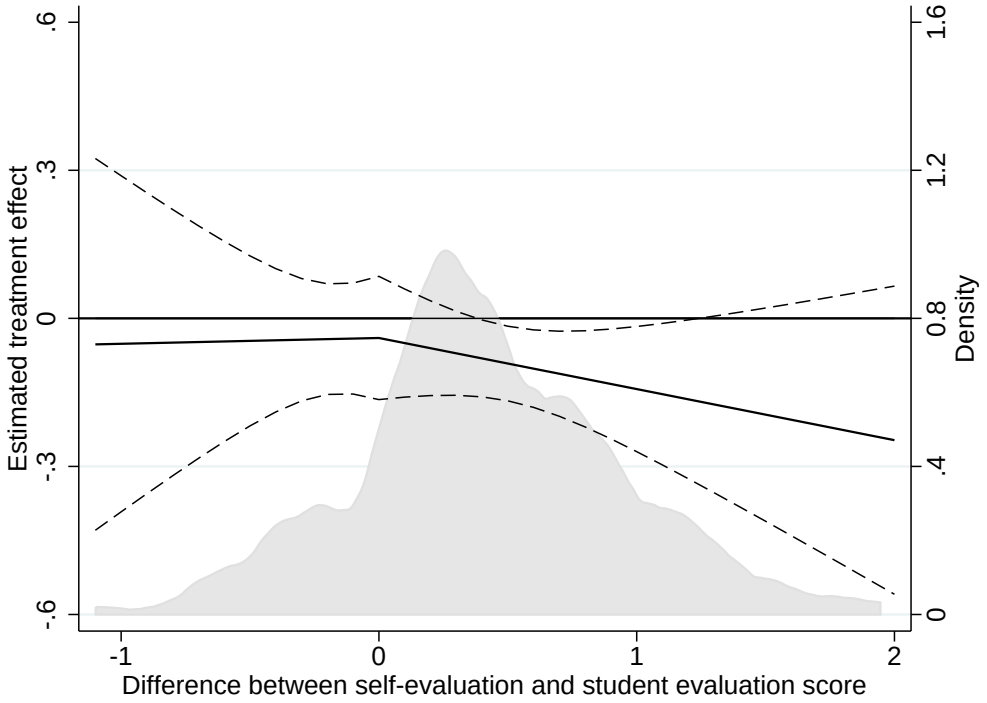
Notes: This figure shows the estimated treatment effect given the difference between a teacher's first-period average self-assessment score and her average first-period student evaluation score ($\Delta self_i$). Dashed lines show the 95% confidence interval. The grey area shows a kernel density of the observations.

Figure 5: Estimated effect of feedback by the difference between a teacher's student evaluation score and her team's average score



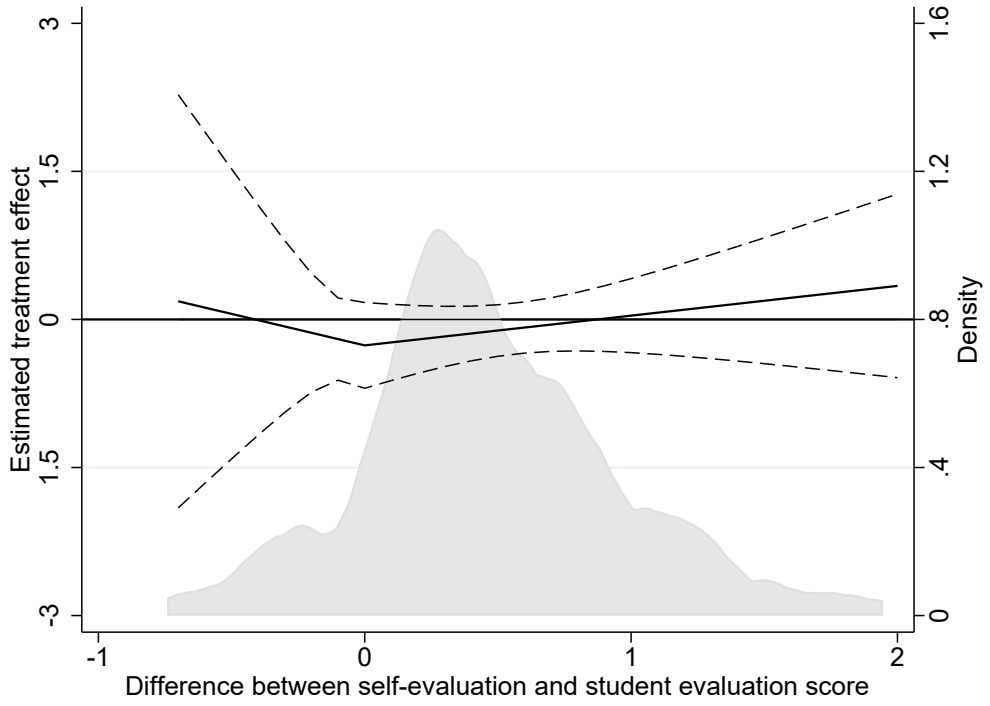
Notes: This figure shows the estimated treatment effect given the difference between a teacher's first-wave average student evaluation score and the average of all first-wave average student evaluation scores of the teachers in her team ($\Delta team_i$). Dashed lines show the 95% confidence interval. The grey area shows a kernel density of the observations.

Figure 6: Estimated effect of feedback on teachers' self-evaluation



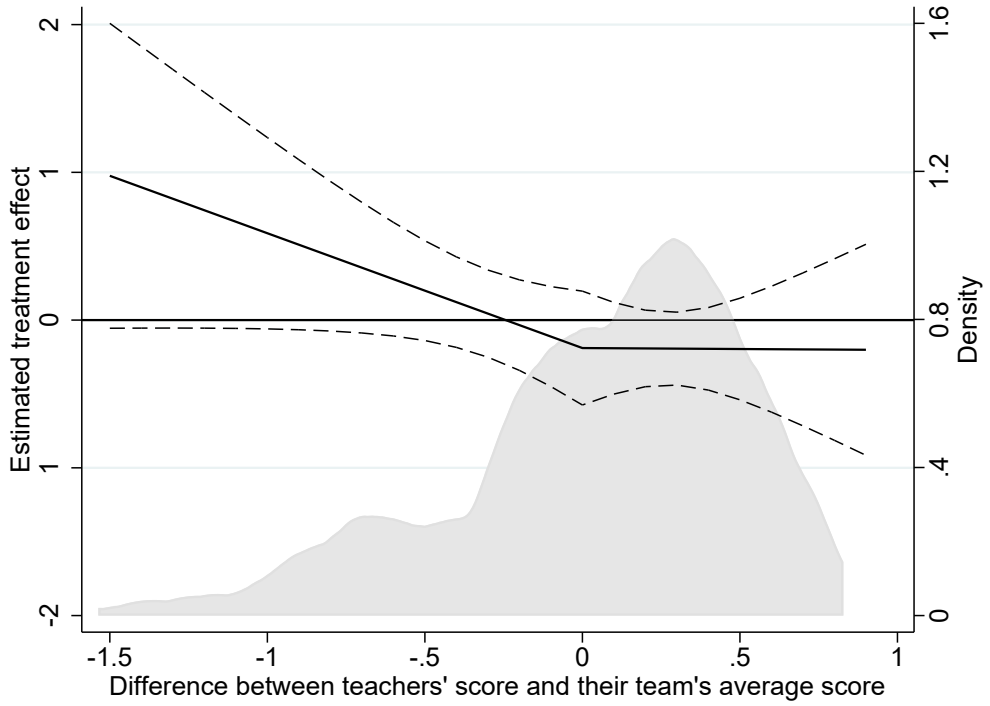
Notes: This figure shows the estimated treatment effect on teachers' average self-evaluation score given the difference between a teacher's first-period average self-evaluation score and her average first-period student evaluation score ($\Delta self_i$). Dashed lines show the 95% confidence interval. The grey area shows a kernel density of the observations.

Figure 7: Estimated effect of feedback on teachers' job satisfaction by the difference between a teacher's self-evaluation score and her student evaluation score



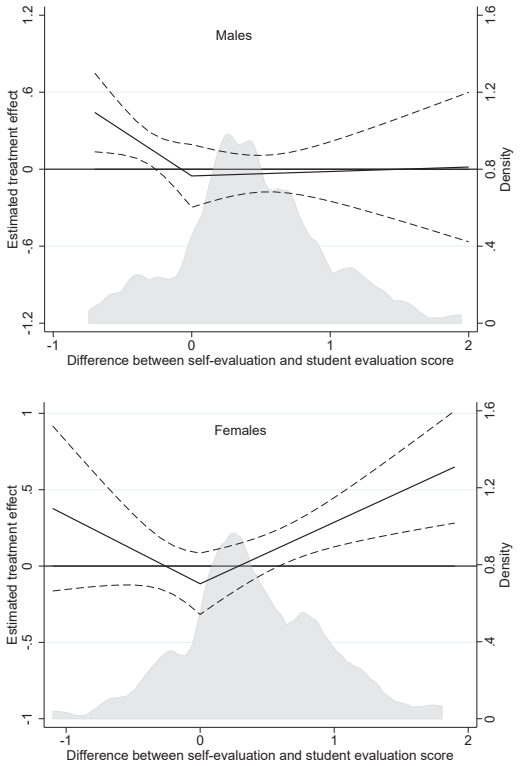
Notes: This figure shows the estimated treatment effect on teachers' job satisfaction given the difference between a teacher's first-period average self-evaluation score and her average first-period student evaluation score ($\Delta self_i$). Dashed lines show the 95% confidence interval. The grey area shows a kernel density of the observations.

Figure 8: Estimated effect of feedback on teachers' job satisfaction by the difference between a teacher's own student evaluation score in the first wave and her team's average score



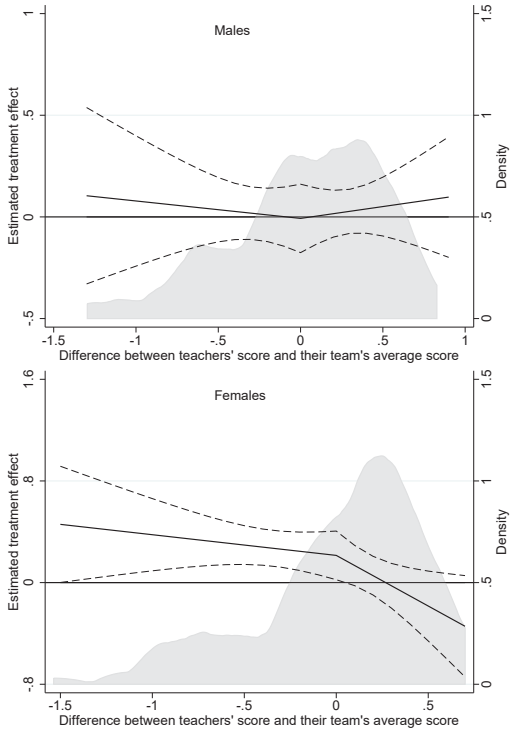
Notes: This figure shows the estimated treatment effect on teachers' job satisfaction given the difference between a teacher's first-wave average student evaluation score and the average of all first-wave average student evaluation scores of the teachers in her team ($\Delta team_i$). Dashed lines show the 95% confidence interval. The grey area shows a kernel density of the observations.

Figure 9: Estimated effect of feedback by the difference between a teacher's self-evaluation score and her student evaluation score, by teachers' gender



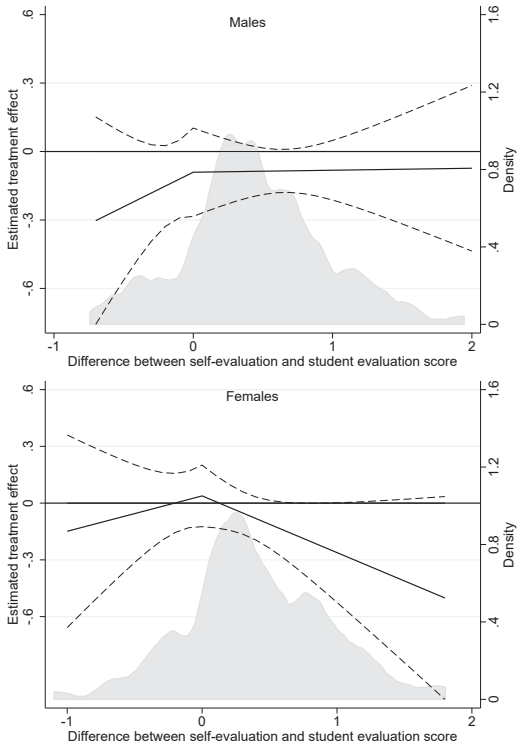
Notes: This figure shows the estimated treatment effect given the difference between a teacher's first-period average self-assessment score and her average first-period student evaluation score ($\Delta self_i$), by teachers' gender. Dashed lines show the 95% confidence interval. The grey areas show kernel densities of the observations.

Figure 10: Estimated effect of feedback by the difference between a teacher's own student evaluation score in the first wave and her team's average score, by teachers' gender



Notes: This figure shows the estimated treatment effect given the difference between a teacher's first-wave average student evaluation score and the average of all first-wave average student evaluation scores of the teachers in her team ($\Delta team_i$), by teachers' gender. Dashed lines show the 95% confidence interval. The grey areas show kernel densities of the observations.

Figure 11: Estimated effect of feedback on teachers' self-evaluation, by teachers' gender



Notes: This figure shows the estimated treatment effect on teachers' average self-evaluation score given the difference between a teacher's first-period average self-evaluation score and her average first-period student evaluation score ($\Delta self_i$), by teachers' gender. Dashed lines show the 95% confidence interval. The grey areas show kernel densities of the observations.

A Appendix

The student evaluation form (translated from Dutch):

		Disagree	Disagree somewhat	Disagree somewhat / Agree somewhat	Agree somewhat	Agree	Do not know /not applicable
1	The teacher is well able to answer questions about the contents of the course	☐	☐	☐	☐	☐	☐
2	The teacher ensures variety in class	☐	☐	☐	☐	☐	☐
3	The teacher keeps me interested	☐	☐	☐	☐	☐	☐
4	The teacher adapts to what I already know and can do	☐	☐	☐	☐	☐	☐
5	The teacher is able to explain the connection to the real world	☐	☐	☐	☐	☐	☐
6	The teacher is well able to answer my questions	☐	☐	☐	☐	☐	☐
7	The teacher makes me feel comfortable in class	☐	☐	☐	☐	☐	☐
8	The teacher is able to maintain order in class	☐	☐	☐	☐	☐	☐
9	The teacher deals with all students in an equal manner	☐	☐	☐	☐	☐	☐
10	The teacher stimulates independent work	☐	☐	☐	☐	☐	☐
11	The teacher is clear in his/her expectations of me	☐	☐	☐	☐	☐	☐
12	The teacher checks whether I did my assignments or homework	☐	☐	☐	☐	☐	☐
13	The teacher is consistent in adhering to appointments	☐	☐	☐	☐	☐	☐
14	The teacher answers e-mails on time (within 2 workdays)	☐	☐	☐	☐	☐	☐
15	The teacher grades assignments within 14 days	☐	☐	☐	☐	☐	☐
16	I am able to have a discussion with my teacher if I disagree with the teacher	☐	☐	☐	☐	☐	☐
17	The teacher takes me seriously	☐	☐	☐	☐	☐	☐
18	The teacher addresses students regarding their behavior	☐	☐	☐	☐	☐	☐
19	The teacher is an example to me in both speech and behavior	☐	☐	☐	☐	☐	☐
20	Comments/recommendations: You can write any comments or recommendations here using up to 300 characters.						

Name teacher: [redacted]

Table A.1: Descriptives questionnaire 2011, by item

		Student evaluations		Self-assessment		Correlation
		mean	st.dev.	mean	st.dev.	student and self-assessment
item	The teacher...					
1	is well able to answer questions	4.27	(0.59)	4.90	(0.42)	0.11
2	ensures variety in class	3.87	(0.71)	4.45	(0.62)	0.09
3	keeps me interested	3.81	(0.74)	4.30	(0.61)	0.21
4	adapts to what I already know and can do	3.91	(0.56)	4.40	(0.67)	-0.02
5	is able to explain the connection to the real world	4.18	(0.58)	4.66	(0.68)	0.21
6	is well able to answer my questions	4.27	(0.53)	4.77	(0.45)	0.04
7	makes me feel comfortable in class	4.32	(0.49)	4.67	(0.54)	0.13
8	is able to maintain order in class	3.99	(0.68)	4.42	(0.74)	0.20
9	deals with all students in an equal manner	4.36	(0.51)	4.80	(0.50)	0.06
10	stimulates independent work	4.24	(0.42)	4.57	(0.65)	-0.16
11	is clear in his/her expectations of me	4.14	(0.50)	4.57	(0.61)	0.12
12	checks whether I did my assignments or homework	3.75	(0.61)	4.04	(1.02)	0.17
13	is consistent in adhering to appointments	4.17	(0.53)	4.61	(0.63)	0.13
14	answers e-mails on time	4.09	(0.54)	4.58	(0.81)	0.23
15	grades assignments within 14 days	4.34	(0.52)	4.71	(0.72)	0.21
16	I am able to have a discussion with my teacher	4.34	(0.53)	4.94	(0.26)	0.19
17	takes me seriously	4.47	(0.42)	4.89	(0.26)	-0.03
18	addresses students regarding their behavior	4.42	(0.39)	4.84	(0.39)	0.08
19	is an example to me in both speech and behavior	3.39	(0.71)	4.37	(0.65)	-0.03

Table A.2: Correlation between questionnaire items 2011

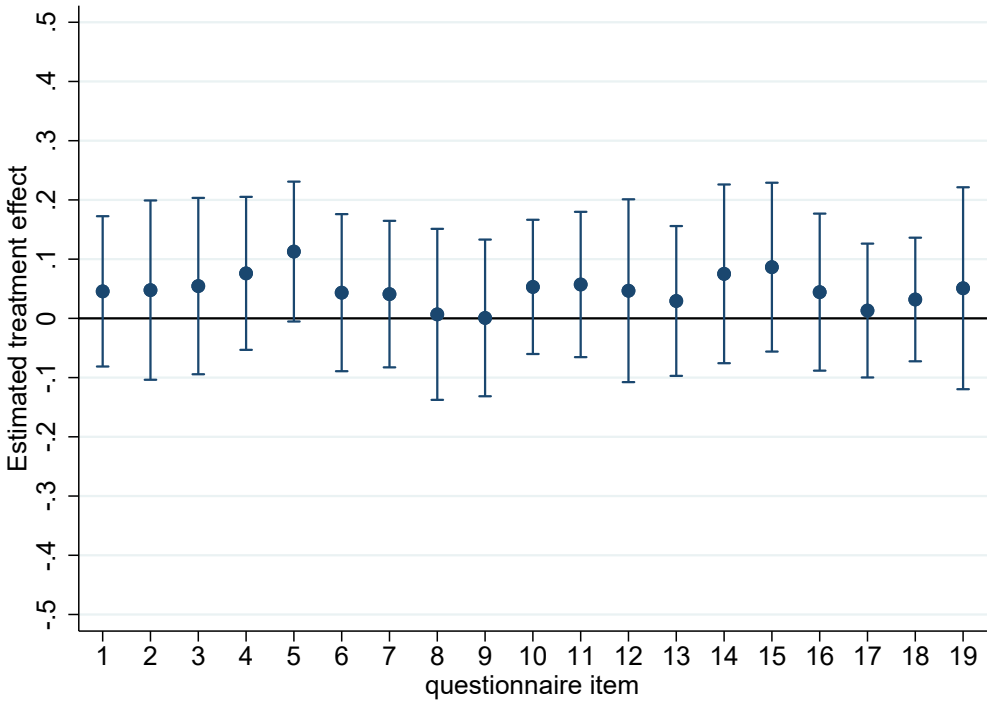
	q1	q2	q3	q4	q5	q6	q7	q8	q9	q10	q11	q12	q13	q14	q15	q16	q17	q18
	student evaluation score																	
q2	0.81																	
q3	0.87	0.90																
q4	0.85	0.85	0.86															
q5	0.90	0.83	0.90	0.87														
q6	0.88	0.82	0.84	0.86	0.84													
q7	0.77	0.77	0.81	0.78	0.80	0.76												
q8	0.81	0.71	0.82	0.75	0.81	0.75	0.72											
q9	0.70	0.69	0.70	0.71	0.71	0.80	0.86	0.69										
q10	0.78	0.75	0.81	0.80	0.82	0.81	0.78	0.75	0.73									
q11	0.83	0.77	0.82	0.80	0.86	0.79	0.71	0.78	0.65	0.81								
q12	0.54	0.61	0.58	0.62	0.60	0.57	0.47	0.48	0.48	0.62	0.71							
q13	0.76	0.72	0.76	0.72	0.79	0.77	0.71	0.71	0.69	0.74	0.83	0.65						
q14	0.58	0.53	0.60	0.64	0.61	0.65	0.60	0.55	0.58	0.65	0.59	0.44	0.64					
q15	0.51	0.47	0.52	0.53	0.57	0.57	0.60	0.44	0.55	0.58	0.60	0.46	0.68	0.66				
q16	0.75	0.76	0.78	0.75	0.84	0.84	0.89	0.66	0.81	0.77	0.66	0.47	0.70	0.59	0.60			
q17	0.75	0.74	0.77	0.75	0.83	0.83	0.89	0.71	0.87	0.77	0.74	0.55	0.74	0.60	0.61	0.88		
q18	0.68	0.67	0.72	0.73	0.66	0.66	0.56	0.75	0.53	0.69	0.80	0.61	0.54	0.54	0.49	0.56	0.60	
q19	0.84	0.83	0.89	0.85	0.84	0.84	0.83	0.80	0.77	0.77	0.81	0.60	0.60	0.60	0.58	0.79	0.80	0.71
	teacher self-assessment																	
q2	0.18																	
q3	0.08	0.29																
q4	0.23	0.37	0.20															
q5	0.24	0.28	0.33	0.19														
q6	0.26	0.14	024	0.29	0.27													
q7	0.31	0.23	0.26	0.34	0.11	0.30												
q8	0.14	0.22	0.32	0.18	0.25	0.20	0.28											
q9	0.13	0.07	0.13	0.13	0.05	0.15	0.33	0.17										
q10	0.27	0.32	0.09	0.30	0.31	0.25	0.28	0.16	0.18									
q11	0.28	0.29	0.23	0.22	0.31	0.24	0.28	0.51	0.09	0.37								
q12	0.21	0.29	0.11	0.11	0.20	0.16	0.27	0.23	0.11	0.48	0.51							
q13	0.15	0.26	0.16	0.32	0.10	0.19	0.31	0.33	0.37	0.37	0.37	0.38						
q14	0.06	-0.06	0.07	0.11	0.00	0.06	0.19	-0.01	0.00	0.14	0.16	0.10	0.13					
q15	0.02	0.02	0.10	0.16	-0.04	0.04	0.10	0.09	-0.09	-0.11	0.09	0.09	0.11	0.26				
q16	0.31	0.10	0.21	0.22	0.32	0.26	0.31	0.13	0.25	0.19	0.11	0.05	0.10	0.02	-0.02			
q17	0.00	0.03	0.04	0.26	0.05	0.23	0.13	0.00	0.27	0.15	0.15	0.01	0.17	-0.01	0.14	0.21		
q18	0.22	0.21	0.17	0.18	0.22	0.22	0.32	0.38	0.17	0.29	0.36	0.24	0.22	0.14	0.14	0.26	0.11	
q19	0.21	0.09	0.05	0.20	0.25	0.18	0.18	-0.02	0.18	0.37	0.23	0.49	0.29	0.11	0.08	0.03	0.21	0.17

Table A.3: The effect of feedback content on attrition (linear probability model)

Dependent variable: drop-out after first year (0 = no; 1 = yes)			
	(1)	(2)	(3)
Treatment	-0.046	-0.008	-0.017
	(0.045)	(0.090)	(0.080)
Δself+		0.027	
		(0.094)	
Δself+ × treatment		-0.096	
		(0.134)	
Δself−		-0.213	
		(0.107**)	
Δself− × treatment		0.623	
		(0.291**)	
Δteam+			0.028
			(0.156)
Δteam+ × treatment			-0.161
			(0.237)
Δteam−			-0.065
			(0.110)
Δteam− × treatment			-0.018
			(0.162)
Constant	-0.091	-0.137	0.111
	(0.188)	(0.226)	(0.196)
Individual controls	Yes	Yes	Yes
Observations	282	198	282
R^2	0.053	0.097	0.062

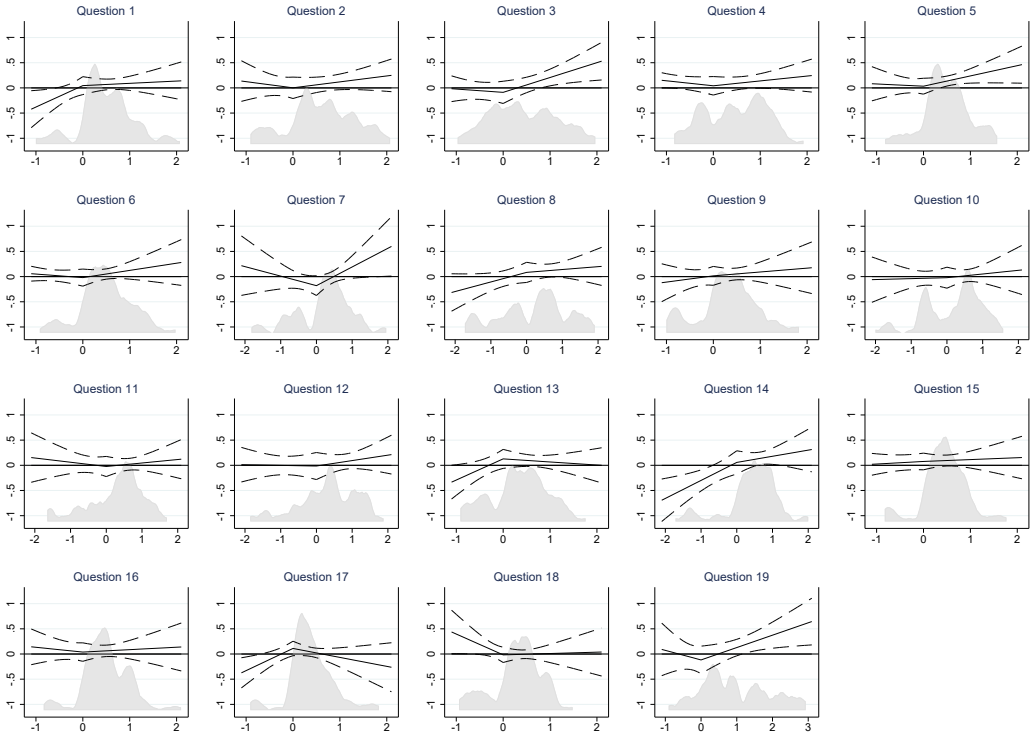
Notes: *, **, and *** indicate significance based on a two-sided test at the .10, .05, and .01 level, respectively. Individual controls are gender, age, tenure, and full-time equivalent. Variable Δself is the difference between a teacher's first-period average self-assessment score and her average first-period student evaluation score. Variable Δteam is the difference between a teacher's first-period average student evaluation score and the average of all first-period average student evaluation scores of the teachers in her team.

Figure A.1: Average treatment effect per questionnaire item



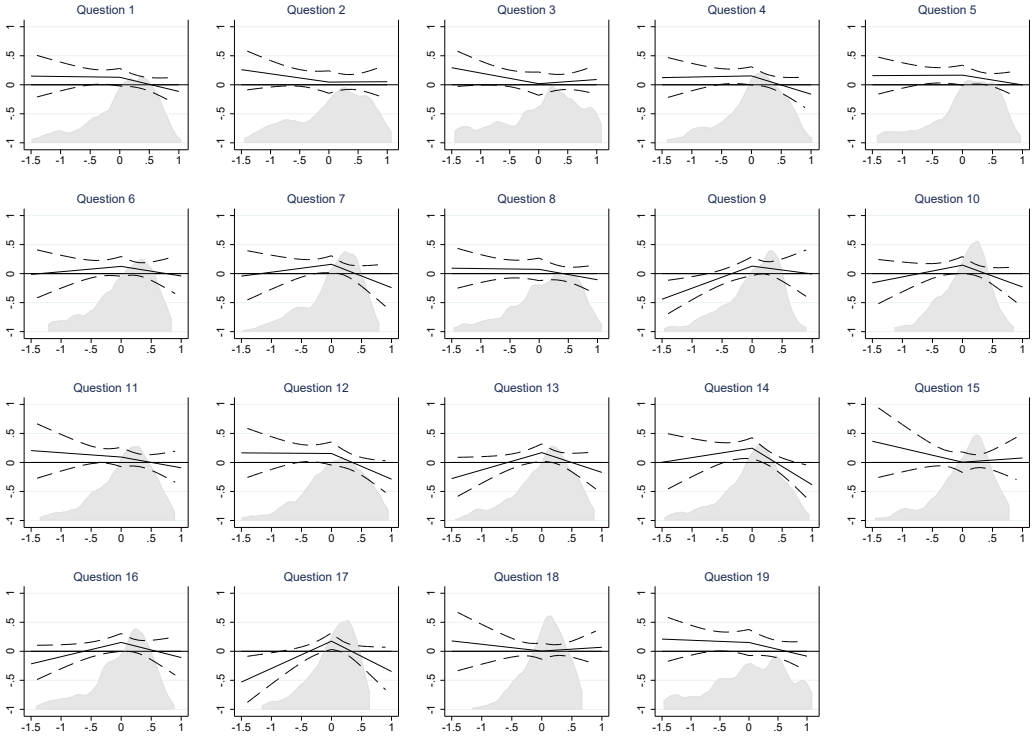
Notes: The dots are the estimated average treatment effect per questionnaire item. The capped lines show the 95% confidence interval of each estimate.

Figure A.2: Estimated effect of feedback by the difference between a teacher's self-evaluation score and her student evaluation score, by questionnaire item



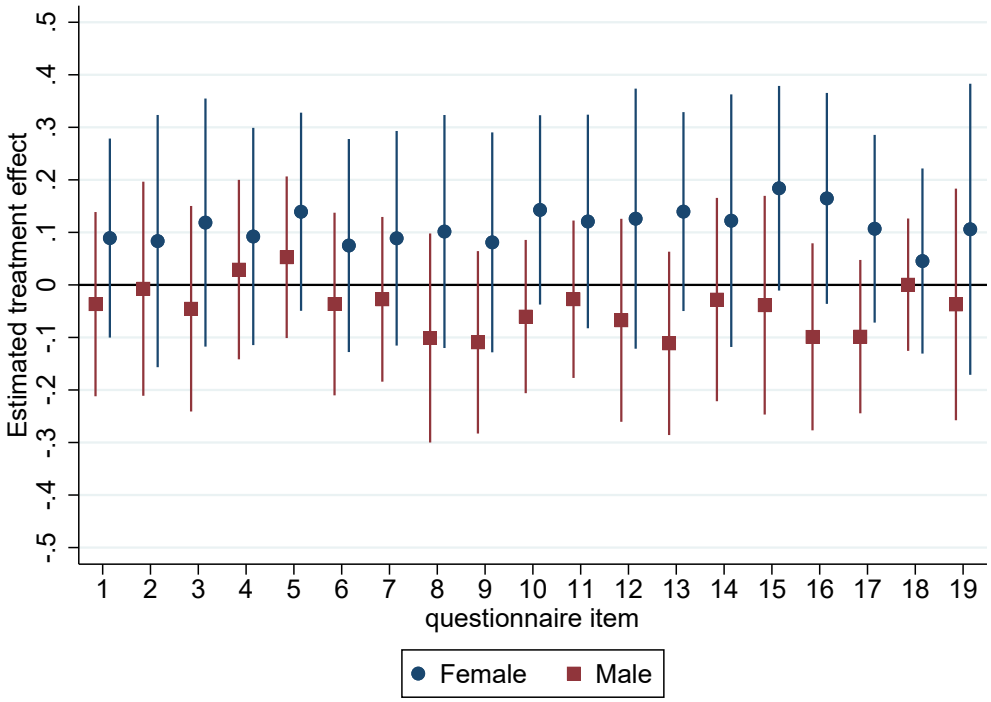
Notes: Each plot in this figure shows, for a given questionnaire item, the estimated treatment effect given the item-specific difference between a teacher's first-period average self-assessment score and her average first-period student evaluation score ($\Delta self_i$). Dashed lines show the 95% confidence interval. The grey area shows a kernel density of the observations.

Figure A.3: Estimated effect of feedback by the difference between a teacher's student evaluation score and her team's average score, by questionnaire item



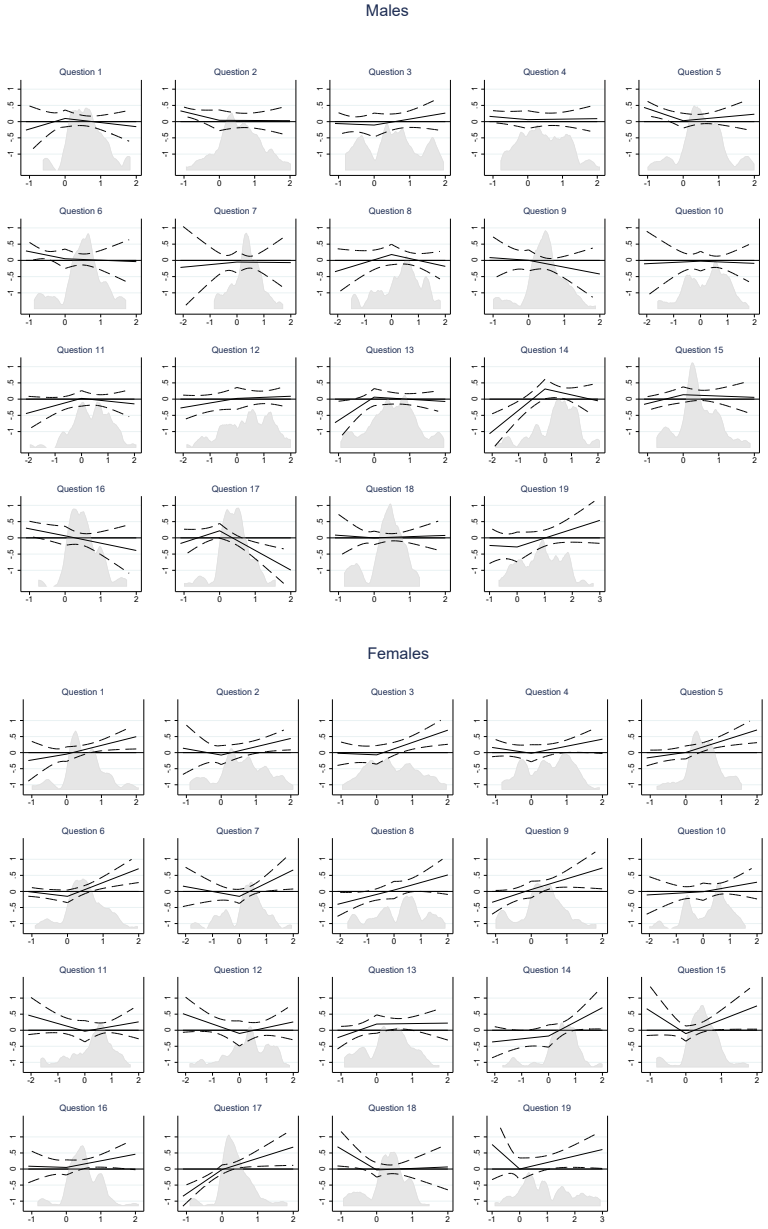
Notes: Each plot in this figure shows, for a given questionnaire item, the estimated treatment effect given the item-specific difference between a teacher's first-wave average student evaluation score and the average of all first-wave average student evaluation scores of the teachers in her team ($\Delta team_i$). Dashed lines show the 95% confidence interval. The grey area shows a kernel density of the observations.

Figure A.4: Average treatment effect per questionnaire item, by gender



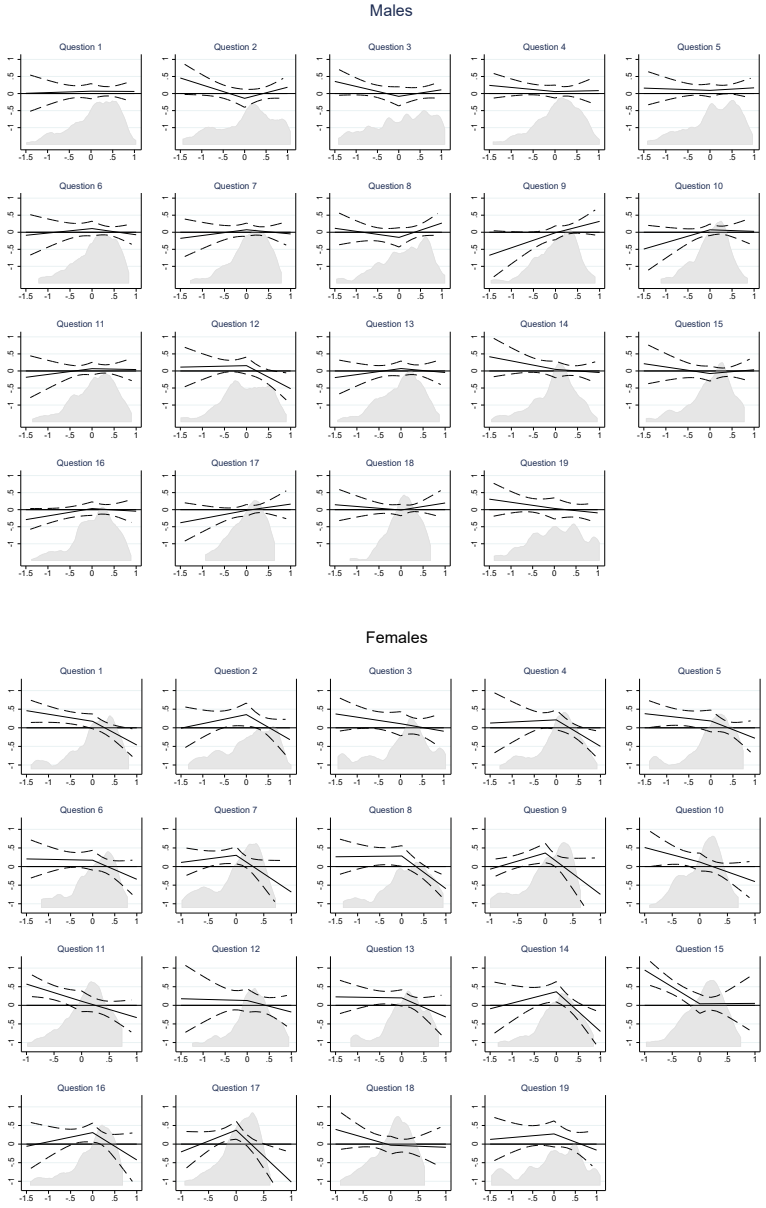
Notes: The dots and squares are the estimated average treatment effect per questionnaire item. The lines show the 95% confidence interval of each estimate.

Figure A.5: Estimated effect of feedback by the difference between a teacher's self-evaluation score and her student evaluation score, by questionnaire item and gender



Notes: Each plot in this figure shows, for a given questionnaire item, the estimated treatment effect given the item-specific difference between a teacher's first-period average self-assessment score and her average first-period student evaluation score ($\Delta self_i$). Dashed lines show the 95% confidence interval. The grey area shows a kernel density of the observations.

Figure A.6: Estimated effect of feedback by the difference between a teacher's student evaluation score and her team's average score, by questionnaire item and gender



Notes: Each plot in this figure shows, for a given questionnaire item, the estimated treatment effect given the item-specific difference between a teacher's first-wave average student evaluation score and the average of all first-wave average student evaluation scores of the teachers in her team ($\Delta team_i$). Dashed lines show the 95% confidence interval. The grey area shows a kernel density of the observations.