

# The Effects of Student Feedback to Teachers: Evidence from a Field Experiment\*

Margaretha Buurman<sup>†</sup>, Josse Delfgaauw<sup>‡</sup>, Robert Dur<sup>§</sup> and Robin Zoutenbier<sup>¶</sup>

September 28, 2018

## Abstract

We conducted a field experiment to examine the effects of student feedback to teachers at a large Dutch school for intermediate vocational education. Students evaluated all teachers, but only a randomly selected group of teachers received feedback. Additionally, we asked all teachers before as well as after the experiment to assess their own performance on the same items. We find a precisely estimated zero average treatment effect of receiving feedback on student evaluation scores a year later. Only those teachers whose self-assessment before the experiment is much more positive than their students' evaluations 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 turn out to be unresponsive to student feedback.

**Keywords:** field experiment, feedback, teachers, student evaluations, gender differences.

**JEL:** C93, I2, M5

---

\*We gratefully acknowledge comments and suggestions by 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 Workshop 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.

<sup>†</sup>Free University Amsterdam. E-mail: m.w.j.m.buurman@vu.nl

<sup>‡</sup>Erasmus University Rotterdam, Tinbergen Institute. E-mail: delfgaauw@ese.eur.nl

<sup>§</sup>Erasmus University Rotterdam, Tinbergen Institute, CESifo, and IZA. E-mail: dur@ese.eur.nl

<sup>¶</sup>Ministry of Finance. E-mail: r.zoutenbier@minfin.nl

# 1 Introduction

Regular provision of performance 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 have investigated the effect of receiving feedback on performance. In a variety of organizations and contexts, these studies have shown that the provision of feedback can have sizeable positive effects on performance (Azmat and Iriberry 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. 2017). Barankay (2012) and Bandiera et al. (2013) show that feedback can also have an adverse effect on performance.

Providing employees with feedback on their performance has also become increasingly popular in education. Many schools use students' evaluations of teachers to enable and motivate teachers to improve teaching. Moreover, evaluations sometimes play a role in tenure, bonus, and promotion decisions (Watts and Becker, 1999). There is an extensive literature that studies the use of students' evaluations in teaching (see for instance Cohen 1980 and Marsh 2007 for overviews of the literature). In general, studies find positive but small effects of students' feedback on the performance of teachers.

This paper studies the effect of students' feedback on the performance of teachers by conducting a field experiment at a large Dutch school for intermediate vocational education. Students were asked to evaluate their teachers using a 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 estimate the effect of receiving feedback on teachers' performance by examining students' evaluations of teachers a year later.<sup>1</sup> In contrast to most previous studies (Centra,

---

<sup>1</sup>There are no standardized test scores or other objective measures of student performance available. 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. (2018) find a positive correlation for male teachers, while there is little correlation for female teachers in their sample. Note also that, even if the correlation in our context would be negative, this need not imply that

1973, being the main exception), we also investigate 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 after the experiment. Another difference with most previous studies is that we examine the effect of feedback over a much longer period of time, namely a full year. Earlier studies are typically restricted to studying the effects of feedback within a semester.

The results of our experiment show that receiving feedback has 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 results differ somewhat from the findings of the existing studies mentioned above. A possible explanation for the lack of a 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 in Azmat et al. (2017) in the context of providing relative performance information to students.

Next, we study whether the content of the feedback matters for the effect of receiving feedback on performance. For that purpose, we compare the student feedback with the prior self-assessment of performance on the same items. We expect that teachers whose self-assessment deviates more from the students’ evaluations respond more strongly, as student feedback may contain more ‘news’ in that case. In line with this, we find 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 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.

---

improvements in student evaluations caused by teachers’ response to students’ feedback lead to a worsening of student performance.

How a teacher’s performance compares to the performance 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 co-workers, 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 perform better than their team. We do find a positive but small (and only marginally significant) effect of feedback for workers who perform worse than their team.

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. 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 Iriberry (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’ performance increases significantly after learning that their student evaluation score falls below their self-assessment score as well as when they learn they perform worse than 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 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 find a similar result when we look at the effect of the content of feedback.

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 the end of 2011 and the beginning of 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 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 Set-up of the experiment**

The experiment is based on two waves of student evaluations of teachers. The first wave took place at the end of 2011, the second wave at the end of 2012. In a pilot prior to 2011, six teams had implemented student evaluation surveys that consisted of 19 items. After analyzing the outcomes of these surveys, we used an adjusted version of this questionnaire in our study. 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 Appendix A. It consists of 19 statements, to which students could respond

on a 5-point scale ranging from ‘disagree’ to ‘agree’,<sup>2</sup> as well as a space for comments and recommendations. The questionnaire includes statements on teacher quality, organizational aspects, and interpersonal skills.

In both years, the questionnaires were administered at the end of the first teaching period in the school year. Before the start of the school year, teachers were informed through an information bulletin that student evaluations would take place. Further, in 2011, teachers were informed 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. In 2012, teachers were informed that all of them 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).

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. All teachers were asked to complete a self-assessment questionnaire on the same items as contained in the student evaluation questionnaire.<sup>3</sup>

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,

---

<sup>2</sup>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.

<sup>3</sup>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.

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.<sup>4</sup> 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 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.<sup>5</sup> 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. Moreover, we obtain balance across teams.

The teachers in the treatment group received their feedback in Spring 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 groups 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 groups did not receive their individual student evaluation scores, but they did receive their self-assessment scores as well as the team scores.<sup>6</sup>

To study the effect of receiving feedback, our main performance measure

---

<sup>4</sup>We excluded statements 14 and 15 here because these consider factual statements regarding time taking for answering e-mails and grading (see Appendix A). We expected that on these items, students' answers were unlikely to surprise 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.

<sup>5</sup>Teachers who did not complete the self-assessment were randomly assigned to treatment and control.

<sup>6</sup>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 stratified assignment to treatment and control within teams 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.



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.

At the end of 2012, we conducted another wave of student evaluations, using the same questionnaire and the same procedure. This time, all teachers were informed that they would receive their 2012 student evaluation scores, which happened in Spring 2013 through e-mail. 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.

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 at the end of 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”.<sup>7</sup>

### 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 above. 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. There are some outliers in the data, but 95% of all teachers in the analysis are evaluated by 10 to 55 students per wave. The number of evaluations per teacher may differ due to differences in class size or differences in response rates across classes. Below, we first provide descriptive statistics for the 242 teachers in the analysis and subsequently discuss attrition.

---

<sup>7</sup>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.

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. 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. On average, teachers' self-assessment score is 4.60, which is considerably higher than the evaluations by their students. There is no significant difference in teachers' self-evaluations between the treatment group and the control group. On observable characteristics, teachers in the treatment and the control 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. Only the differences in working hours and tenure are marginally significant at the 10-percent level.<sup>8</sup>

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.

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.<sup>9</sup> 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

---

<sup>8</sup>We discuss the differences between male and female teachers in Section 6.

<sup>9</sup>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.

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.<sup>10</sup>

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 Appendix A) averaged over all students who evaluate the teacher in a given year.<sup>11</sup> 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 time-fixed effects and teacher-fixed effects, denoted by  $\lambda_t$  and  $\theta_i$ , respectively. The regression equation reads:

$$y_{it} = \gamma T_{it} + \lambda_t + \theta_i + \varepsilon_{it}. \quad (1)$$

The estimated average treatment effect of receiving feedback is given by  $\gamma$ . In all our estimations, we cluster standard errors at the teacher level.<sup>12</sup>

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  $\Delta self_i$ , which denotes

---

<sup>10</sup>We further examine the issue of selective attrition in Section 5.

<sup>11</sup>Using instead the average score excluding statements 14 and 15 (as used to stratify assignment to treatment) does not affect our results in any important way.

<sup>12</sup>Equation (1) is specified at the teacher level. We also estimate the average treatment effect at the student level.

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:

$$y_{it} = \gamma T_{it} + \varphi(T_{it} \times \Delta self_i) + \psi(E_t \times \Delta self_i) + \lambda_t + \theta_i + \varepsilon_{it}. \quad (2)$$

Note that we also interact  $\Delta self_i$  with dummy variable  $E_t$  that takes value 1 in the second year of our experiment and is zero otherwise. This interaction accounts for correlations between second-year evaluation scores and  $\Delta self_i$  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:

$$y_{it} = \gamma T_{it} + \varphi_p(T_{it} \times \Delta self_i^{pos}) + \varphi_n(T_{it} \times \Delta self_i^{neg}) + \psi_p(E_t \times \Delta self_i^{pos}) + \psi_n(E_t \times \Delta self_i^{neg}) + \lambda_t + \theta_i + \varepsilon_{it}, \quad (3)$$

where  $\Delta self_i^{pos} = \Delta self_i$  if  $\Delta self_i > 0$  and  $\Delta self_i^{pos} = 0$  if  $\Delta self_i \leq 0$ . Variable  $\Delta self_i^{neg}$  correspondingly captures the negative values of  $\Delta self_i$ .<sup>13</sup>

Second, in a similar way we include the interaction between the treatment dummy and the variable  $\Delta team_i$ , which gives the difference between teacher  $i$ 's average student evaluation score in the first year and the average of the first-year evaluations of all teachers in her team. Hence,  $\Delta team_i$  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 similar regressions using teachers' second-year average self-assessment scores and job satisfaction as dependent variables.

## 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. This effect is quite precisely estimated, with a standard error equal to 0.054 and a 95 percent confidence interval that runs

---

<sup>13</sup>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  $[-0.5, 0.5]$ .

from  $-0.063$  to  $0.149$ . 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 may be that previous studies 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. (2017) 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).

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 difference between the two estimates indicates that the average treatment effect is slightly higher for teachers who are evaluated by relatively few students.<sup>14</sup> 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.<sup>15</sup>

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). The interaction effect is positive but insignificant. The estimated treatment effect for teachers who learn that their first-period self-assessment is equal to their students' assessment is very close to zero at  $0.014$ . For teachers who learn that their students' evaluation score is one point lower than their self-assessment, the treatment effect is  $0.104$  higher. In column 2, we report the results of estimating (3). Figure 4 depicts the estimated effects of receiving feedback. We find that teachers whose own assessment corresponds to students' assessment do not respond to receiving feedback. The estimated treatment effect is positive for teachers who learn that their student evaluation score differs widely from their self-assessment. This effect is significant at the 5-percent level for teachers whose

---

<sup>14</sup>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.

<sup>15</sup>Additionally, we have also estimated the average treatment effect on each of the 19 items of the questionnaire separately. Estimated effects range from  $0.00$  to  $0.11$ , and is significant (at the  $0.06$  level) only for item 5 ("The teacher is able to explain the connection to the real world.").

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).<sup>16</sup>

Second, we examine whether the effect of feedback depends on the gap between a teacher’s first-period student evaluation score and the average score in his team. The third column of Table 5 gives the results of estimating (2) with  $\Delta team_i$  instead of  $\Delta self_i$ . 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).

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. 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 probit estimations on the set of teachers with student evaluation scores in 2011, with a dummy that takes value 1 if a teacher drops out as the dependent variable. As reported in Table A.1 in the Appendix, the estimations show that neither receiving feedback nor the content of this feedback significantly affects the probability of dropping out.<sup>17</sup>

We have seen that on average, teachers’ self-assessment is much more fa-

---

<sup>16</sup>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  $\Delta self_i$ . 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  $\Delta self_i$  in one single regression gives rise to problems of multicollinearity. The correlation between first-period student evaluation score and  $\Delta self_i$  is  $-0.71$ .

<sup>17</sup>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.

vorable 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 ( $\Delta self_i$ ). As expected, the interaction effect is negative, but statistically insignificant. In column 3, we allow the interaction effect to differ for positive and negative values of  $\Delta self_i$ . 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.

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. 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  $\Delta self_i$ . Surprisingly, the estimated interaction effect is positive, but insignificant. Column 3 estimates the relation separately for positive and negative values of  $\Delta self_i$ . 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 inter-

action 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.

## 6 Gender differences in response to feedback

In this section, we analyse 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.<sup>18</sup> 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 fte, 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.<sup>19</sup> 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.

---

<sup>18</sup>Our data do not allow us to examine gender bias in student evaluations. Recently, Boring (2017) and Mengel et al. (2018) 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.

<sup>19</sup>In the self-evaluation, female teachers do rate themselves significantly higher on item 12, but not on the other items.



The regression results can be found in Table 8, and are depicted in Figures 9 to 11. 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  $\Delta self^{pos}$  and  $\Delta self^{neg}$ .<sup>20</sup> As depicted in Figure 9, our finding that receiving ‘bad news’ leads to higher performance 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. We obtain a similar result when replacing  $\Delta self$  with  $\Delta team$  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. Hence, in short, whereas male teachers ignore the feedback provided, female teachers do respond depending on the content of feedback.

## 7 Conclusion

This paper has studied the effects of receiving students’ feedback on teacher performance as measured by student evaluations one year later. We find that on average, receiving students’ feedback has no effect on teacher performance. This contrasts with recent studies on short-run effects of performance feedback, which tend to find positive effects. Our study suggests that effects of feedback (if they exist in our context) are short-lived. A possible remedy 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.

Additionally, we examined whether the response to feedback depends on

---

<sup>20</sup>We find qualitatively similar gender differences when we control for the interaction between the content of feedback and other observable characteristics (age, tenure, and fte). Of course, it is possible that the gender differences are (partially) driven by non-observed factors, leading to omitted variable bias.

the content of feedback. We found that teachers who learn that their students' assessment is much less favorable than their own assessment improve performance 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 Iriberry (2017), What You Know Can't Hurt You? A Field Experiment on Relative Feedback Performance, Mimeo.
- [3] Azmat, Ghazala, and Nagore Iriberry (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 Iriberry (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] 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.
- [16] 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.
- [17] Centra, John A. (1973), Effectiveness of Student Feedback in Modifying College Instruction, *Journal of Educational Psychology*, 65(3): 395-401.
- [18] 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.
- [19] Delfgaauw, Josse, and Robert Dur (2008), Incentives and Workers' Motivation in the Public Sector, *Economic Journal*, 118: 171-191.
- [20] 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.
- [21] 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.
- [22] 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.
- [23] Kuhnen, Camelia M., and Agnieszka Tymula (2012), Feedback, Self-Esteem and Performance in Organizations, *Management Science*, 58(1): 94-113.

- [24] 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.
- [25] Mengel, Friederike, Jan Sauermann, and Ulf Zölitz (2018), Gender Bias in Teaching Evaluations, *Journal of the European Economic Association*, forthcoming.
- [26] Möbius, Markus M., Muriel Niederle, Paul Niehaus, and Tanya S. Rosenblat (2007), Gender Differences in Incorporating Performance Feedback, Mimeo, Harvard University.
- [27] Moldovanu, Benny, Aner Sela, and Xianwen Shi (2007), Contests for Status, *Journal of Political Economy*, 115(2): 338-363.
- [28] 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.
- [29] Ryan, James J., James A. Anderson, and Allen B. Birchler (1980), Student Evaluation: The Faculty Responds, *Research in Higher Education*, 12(4): 317-333.
- [30] 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.
- [31] Suurmond, Guido, Otto Swank, and Bauke Visser (2004), On the Bad Reputation of Reputational Concerns, *Journal of Public Economics*, 88(12): 2817-2838.
- [32] Tran, Anh, and Richard Zeckhauser (2012), Rank as an Inherent Incentive: Evidence from a Field Experiment, *Journal of Public Economics*, 96(9): 645-650.
- [33] 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 <sup>a</sup>					
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 <sup>++</sup>
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 <sup>+++</sup>
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 <sup>++</sup>
Standard deviation	(10.42)	(10.01)	(10.26)	(10.17)	(10.11)
Number of teachers	116	126	242	123	112

Notes: <sup>a</sup> 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 <sup>a</sup>				
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: <sup>a</sup> 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 significant.

Table 3: Descriptive statistics by self-evaluation

	Both self-evaluations completed	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)
Only 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 the groups at the .10, .05, and .01 level, respectively (based on an F-test).



Table 4: Effect of feedback on teachers' performance

Dependent variable: average student evaluation		
	(1) Teacher level	(2) Student level
Treatment	0.043 (0.054)	0.021 (0.046)
Year-fixed effects	Yes	Yes
Teacher-fixed effects	Yes	Yes
Observations	484	15194
Teachers	242	242
Within R <sup>2</sup>	0.016	0.000

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 (0.066)	-0.067 (0.091)	0.062 (0.048)	0.105 (0.076)
$\Delta\text{self} \times \text{treatment}$	0.104 (0.110)			
$\Delta\text{self+} \times \text{treatment}$		0.207 (0.140)		
$\Delta\text{self-} \times \text{treatment}$		-0.275 (0.231)		
$\Delta\text{team} \times \text{treatment}$			-0.090 (0.101)	
$\Delta\text{team+} \times \text{treatment}$				-0.227 (0.225)
$\Delta\text{team-} \times \text{treatment}$				0.001 (0.167)
$\Delta\text{self} \times \text{second period}$	0.216** (0.087)			
$\Delta\text{self+} \times \text{second period}$		0.294** (0.115)		
$\Delta\text{self-} \times \text{second period}$		-0.105 (0.168)		
$\Delta\text{team} \times \text{second period}$			-0.369*** (0.075)	
$\Delta\text{team+} \times \text{second period}$				-0.314** (0.154)
$\Delta\text{team-} \times \text{second period}$				-0.407*** (0.134)
Teacher fixed effects	Yes	Yes	Yes	Yes
Time fixed effect	Yes	Yes	Yes	Yes
Observations	332	332	484	484
Teachers	166	166	242	242
Within R <sup>2</sup>	.166	.218	.217	.218

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  $\Delta\text{self}$  is the difference between a teacher's first-period average self-assessment score and her first-period average student evaluation score. Variable  $\Delta\text{team}$  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 (0.046)	-0.042 (0.059)	-0.040 (0.076)
$\Delta\text{self} \times \text{treatment}$		-0.108 (0.097)	
$\Delta\text{self}+ \times \text{treatment}$			-0.104 (0.122)
$\Delta\text{self}- \times \text{treatment}$			0.012 (0.230)
$\Delta\text{self} \times \text{second period}$		-0.091 (0.060)	
$\Delta\text{self}+ \times \text{second period}$			-0.014 (0.075)
$\Delta\text{self}- \times \text{second period}$			-0.418*** (0.120)
Teacher fixed effects	Yes	Yes	Yes
Time fixed effect	Yes	Yes	Yes
Observations	264	264	264
Teachers	132	132	132
Within $R^2$	.016	.095	.121

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  $\Delta\text{self}$  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 (0.133)	-0.257 (0.214)	-0.263 (0.262)	-0.047 (0.134)	-0.189 (0.233)
$\Delta\text{self}$		-0.407 (0.262)			
$\Delta\text{self}+$			-0.423 (0.278)		
$\Delta\text{self}-$			0.884 (1.789)		
$\Delta\text{self} \times \text{treatment}$		0.292 (0.321)			
$\Delta\text{self}+ \times \text{treatment}$			0.301 (0.375)		
$\Delta\text{self}- \times \text{treatment}$			-0.638 (1.931)		
$\Delta\text{team}$				0.274 (0.211)	
$\Delta\text{team}+$					-0.172 (0.446)
$\Delta\text{team}-$					0.544 (0.361)
$\Delta\text{team} \times \text{treatment}$				-0.419 (0.283)	
$\Delta\text{team}+ \times \text{treatment}$					-0.013 (0.650)
$\Delta\text{team}- \times \text{treatment}$					-0.777 (0.498)
Individual controls	Yes	Yes	Yes	Yes	Yes
Observations	162	130	130	162	162
R <sup>2</sup>	.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  $\Delta\text{self}$  is the difference between a teacher's first-period average self-assessment score and her first-period average student evaluation score. Variable  $\Delta\text{team}$  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.

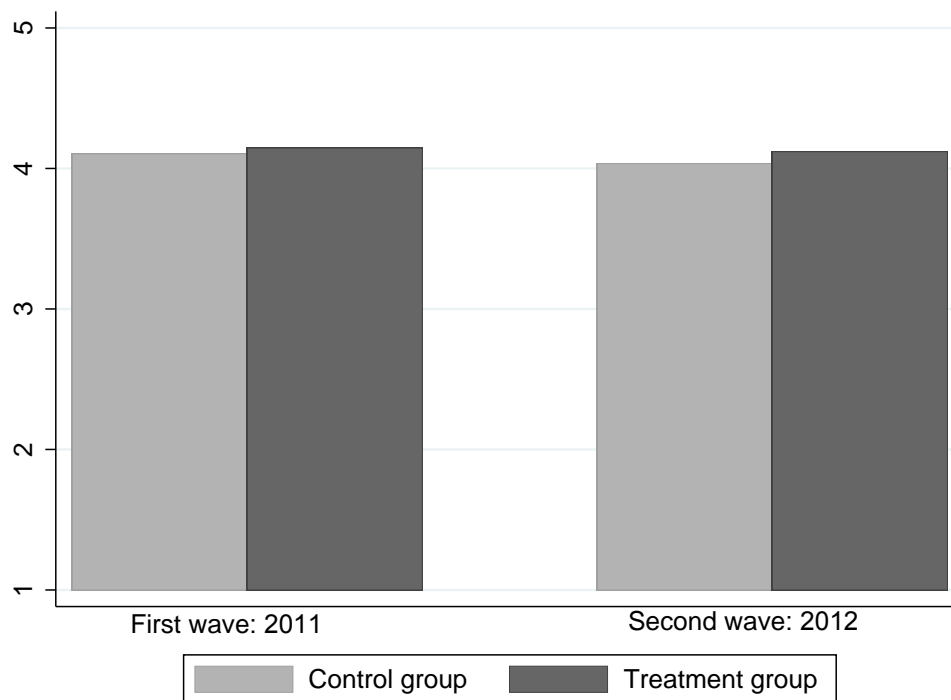
Table 8: Gender differences in the effects of feedback

Dependent variable:	avg student evaluation		avg student evaluation		avg student evaluation		self-evaluation	
	Male (1)	Female (2)	Male (3)	Female (4)	Male (5)	Female (6)	Male (7)	Female (8)
Treatment	-0.043 (0.067)	0.099 (0.088)	-0.053 (0.147)	-0.116 (0.121)	-0.007 (0.102)	0.214* (0.115)	-0.078 (0.110)	0.038 (0.098)
$\Delta\text{self+} \times \text{treatment}$			0.035 (0.230)	0.403** (0.161)			-0.008 (0.144)	-0.230 (0.216)
$\Delta\text{self-} \times \text{treatment}$			-0.707* (0.420)	-0.448 (0.349)			0.341 (0.440)	0.187 (0.333)
$\Delta\text{team+} \times \text{treatment}$					0.116 (0.271)	-0.797* (0.420)		
$\Delta\text{team-} \times \text{treatment}$					-0.086 (0.243)	-0.163 (0.227)		
$\Delta\text{self+} \times \text{second period}$			0.312* (0.176)	0.194 (0.132)			-0.051 (0.099)	0.058 (0.133)
$\Delta\text{self-} \times \text{second period}$			0.561 (0.384)	-0.185 (0.180)			-0.862** (0.403)	-0.340*** (0.092)
$\Delta\text{team+} \times \text{second period}$					-0.377* (0.195)	-0.344 (0.260)		
$\Delta\text{team-} \times \text{second period}$					-0.454*** (0.159)	-0.097 (0.182)		
Teacher fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time fixed effect	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
Within R <sup>2</sup>	.004	.037	.181	.303	.275	.233	0.203	0.136

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  $\Delta\text{self}$  is the difference between a teacher's first-period average self-assessment score and her first-period average student evaluation score. Variable  $\Delta\text{team}$  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.

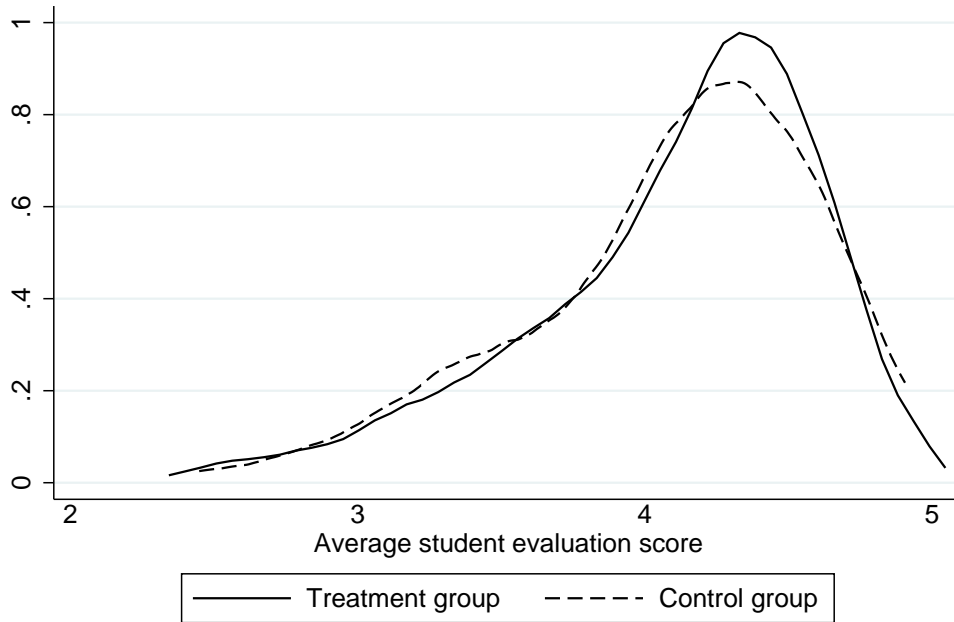
## Figures

Figure 1: Average student evaluation scores by year



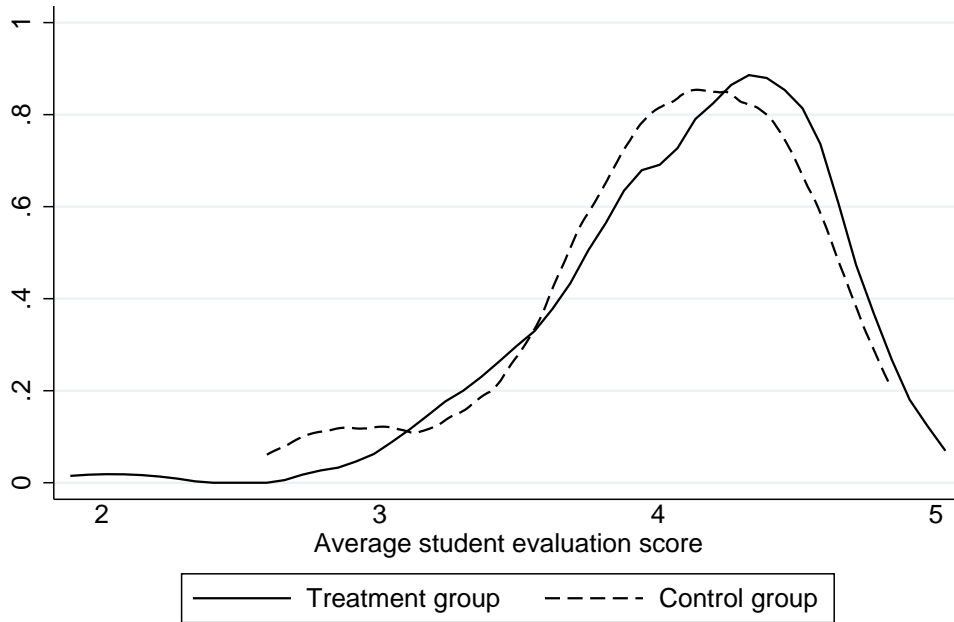
Notes: A student's evaluation of a teacher is defined as the average score on 19 statements on the teacher's performance (see Appendix A). The answer categories for each statement are [1] Disagree, [2] Disagree somewhat, [3] Disagree somewhat/Agree somewhat, [4] Agree somewhat, and [5] Agree.

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 Appendix A). The answer categories for each statement are [1] Disagree, [2] Disagree somewhat, [3] Disagree somewhat/Agree somewhat, [4] Agree somewhat, and [5] Agree.

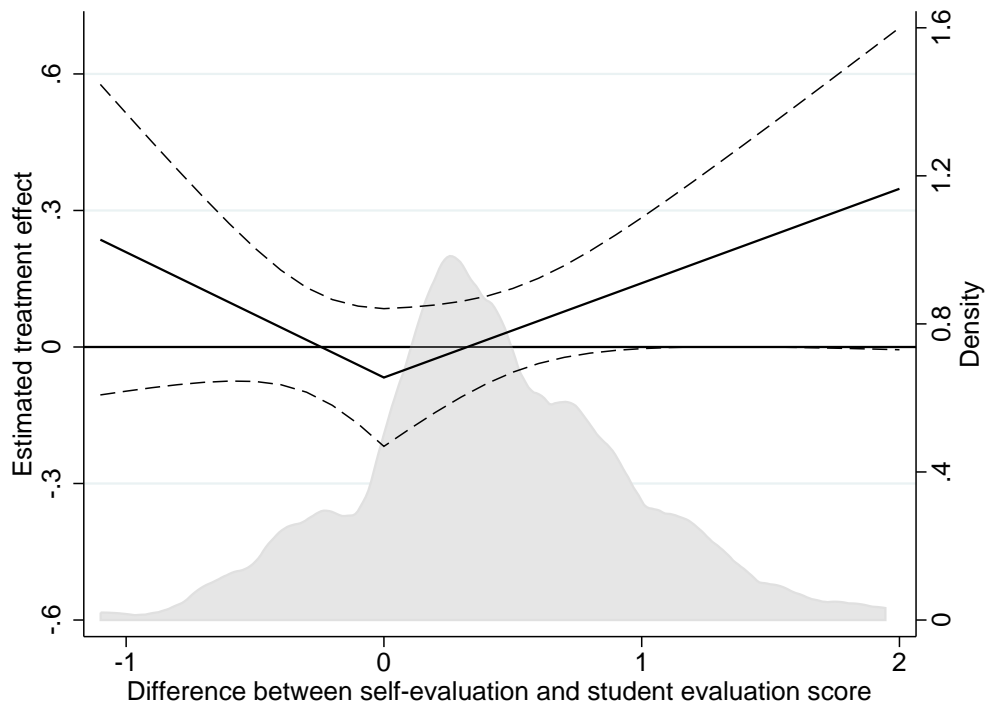
Figure 3: Distribution of student evaluations in the second wave



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 Appendix A). The answer categories for each statement are [1] Disagree, [2] Disagree somewhat, [3] Disagree somewhat/Agree somewhat, [4] Agree somewhat, and [5] Agree.

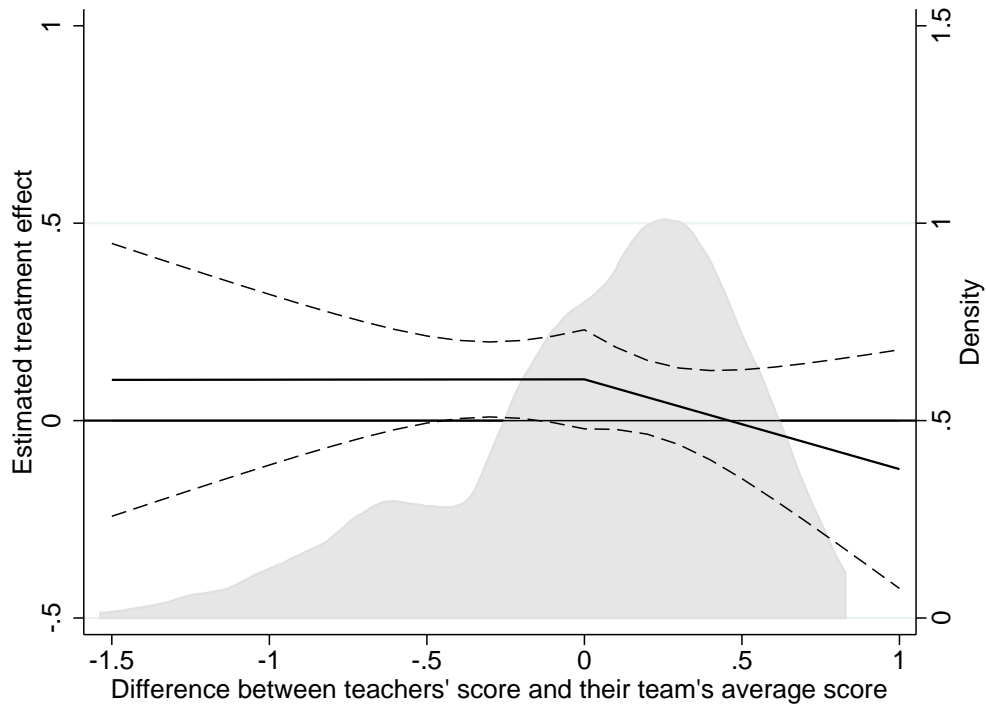


Figure 4: Estimated effect of feedback by the difference between a teacher's self-evaluation score and her student evaluation score



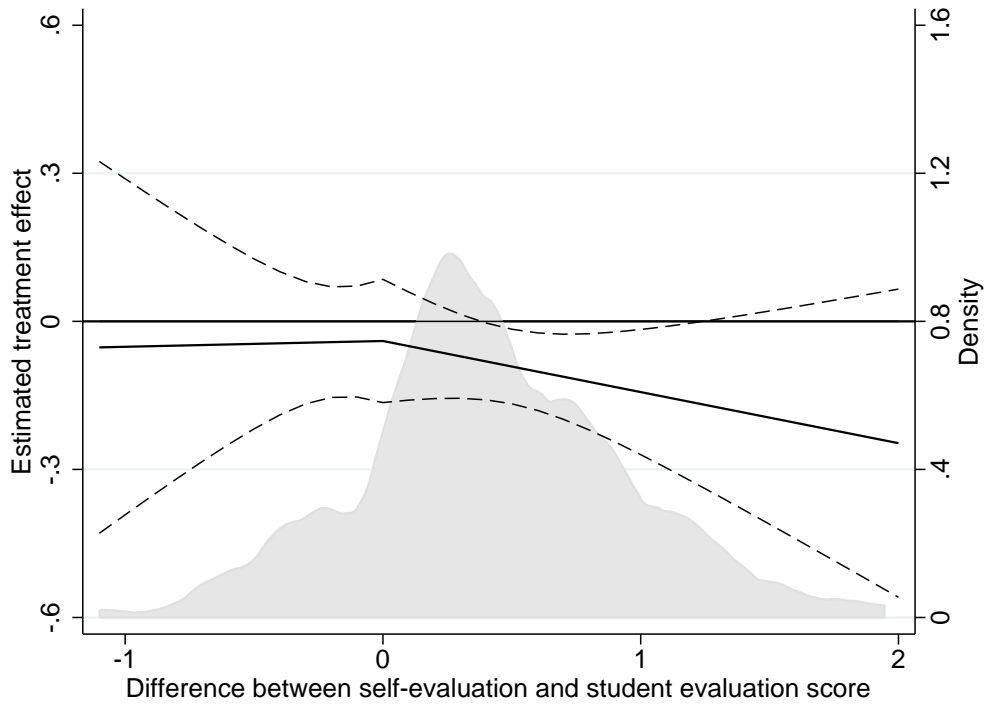
Notes: This figure shows the estimated treatment effect given the difference between a teacher's first-period average self-assessment score and her average first-period student evaluation score ( $\Delta self_i$ ). Dashed lines show the 95% confidence interval. The grey area shows a kernel density of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

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



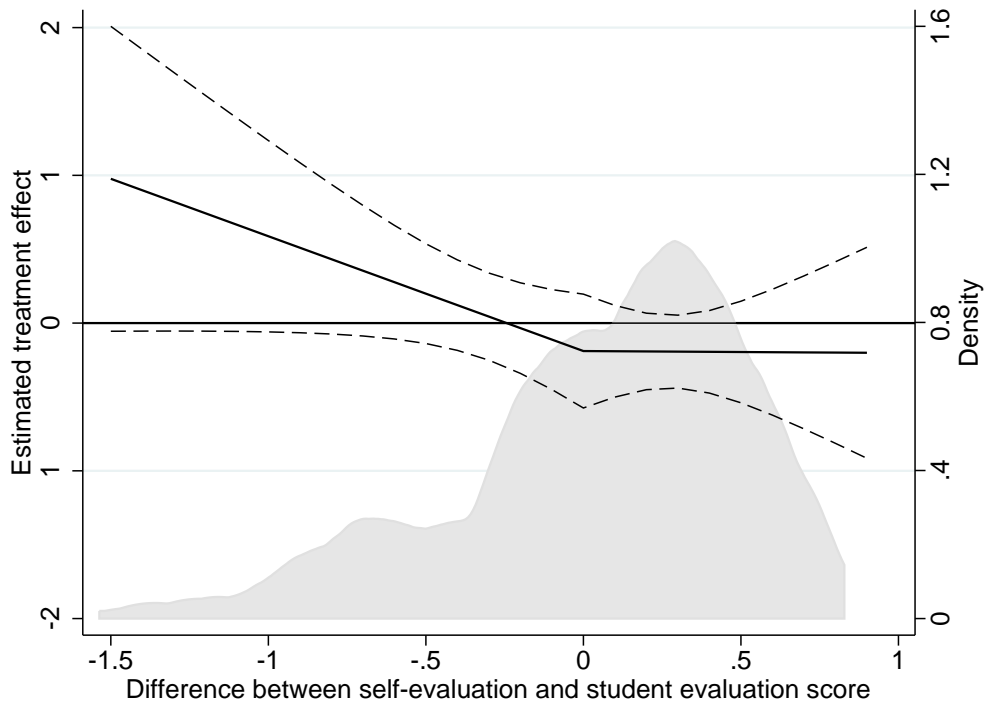
Notes: This figure shows the estimated treatment effect given the difference between a teacher's first-wave average student evaluation score and the average of all first-wave average student evaluation scores of the teachers in her team ( $\Delta team_i$ ). Dashed lines show the 95% confidence interval.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

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



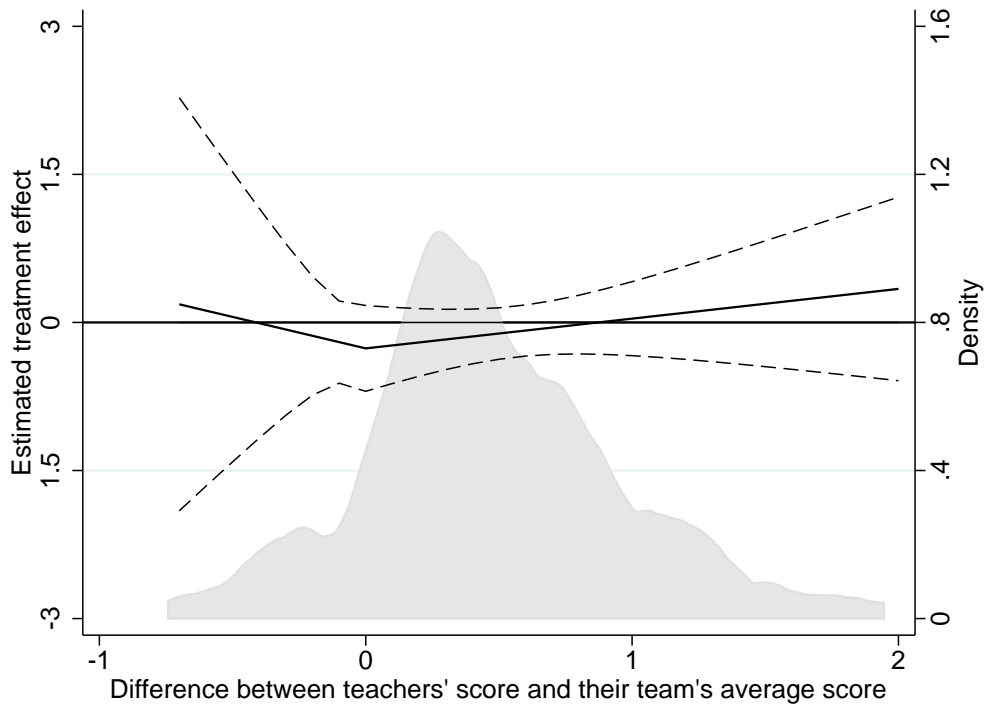
Notes: This figure shows the estimated treatment effect on teachers' average self-evaluation score given the difference between a teacher's first-period average self-evaluation score and her average first-period student evaluation score ( $\Delta self_i$ ). Dashed lines show the 95% confidence interval.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



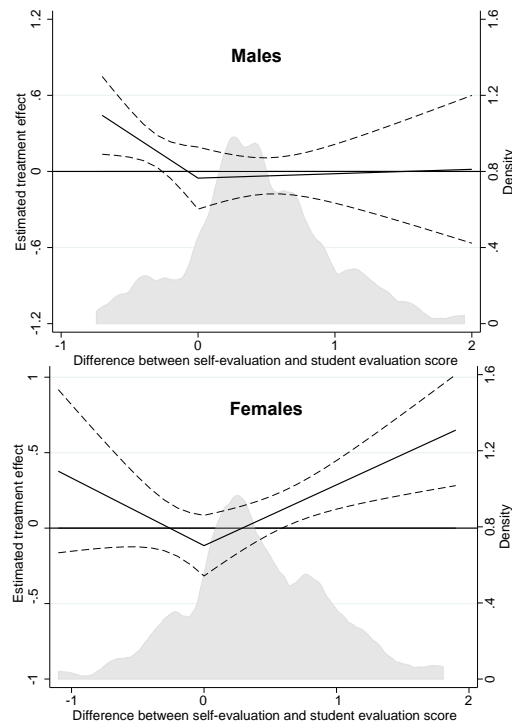
Notes: This figure shows the estimated treatment effect on teachers' job satisfaction given the difference between a teacher's first-period average self-evaluation score and her average first-period student evaluation score ( $\Delta self_i$ ). Dashed lines show the 95% confidence interval. The grey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



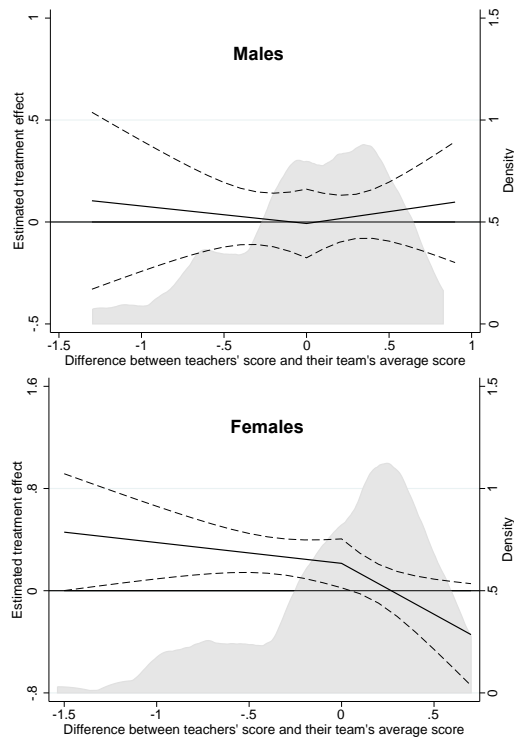
Notes: This figure shows the estimated treatment effect on teachers' job satisfaction given the difference between a teacher's first-wave average student evaluation score and the average of all first-wave average student evaluation scores of the teachers in her team ( $\Delta team_i$ ). Dashed lines 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



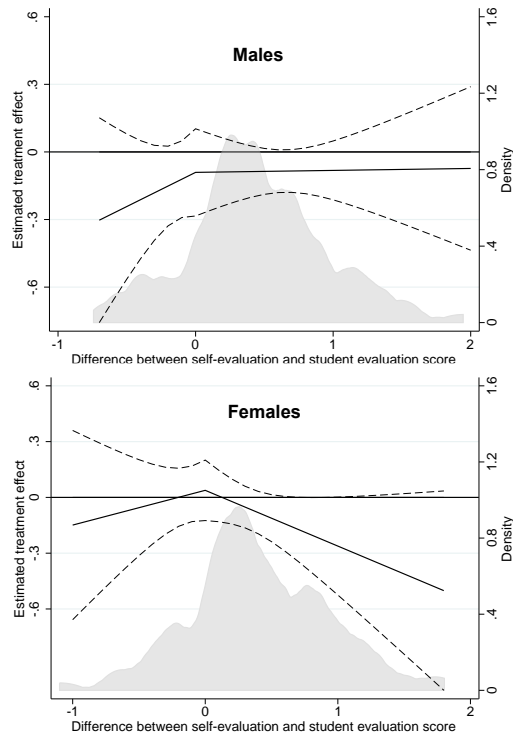
Notes: This figure shows the estimated treatment effect given the difference between a teacher's first-period average self-assessment score and her average first-period student evaluation score ( $\Delta self_i$ ), by teachers' gender. Dashed lines show the 95% confidence interval. The grey areas show 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	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
10	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
11	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
12	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
13	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
14	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
15	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
16	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
17	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
18	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
19	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
20	Comments/recommendations: You can write any comments or recommendations here using up to 300 characters.					

Name teacher: [REDACTED]

Table A.1: The effect of feedback content on attrition (probit estimates)

Dependent variable: drop-out after first year (0 = no; 1 = yes)			
	(1)	(2)	(3)
Treatment	-0.178 (0.188)	-0.490 (0.294)	-0.188 (0.191)
$\Delta_{\text{self}}$		-0.236 (0.240)	
$\Delta_{\text{self}} \times \text{treatment}$		0.356 (0.414)	
$\Delta_{\text{team}}$			-0.110 (0.258)
$\Delta_{\text{team}} \times \text{treatment}$			-0.310 (0.384)
Constant	-2.530 (0.866)***	-1.484 (0.989)	-2.493 (0.870)***
Individual controls	Yes	Yes	Yes
Observations	282	198	282
Pseudo R <sup>2</sup>	0.067	0.081	0.075

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  $\Delta_{\text{self}}$  is the difference between a teacher's first-period average self-assessment score and her average first-period student evaluation score. Variable  $\Delta_{\text{team}}$  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.