Learning from Praise: Evidence from a Field Experiment with Teachers

Maria Cotofan [*]

Abstract

Financial incentive programs for teachers are increasingly common, but little is known about the effectiveness of non-monetary incentives in improving educational outcomes. This field experiment measures how repeated public praise for the best teachers impacts student performance. In treated schools, the students of praised teachers perform better on standardized exams undertaken six months after the intervention. Praised teachers also assign higher marks to their students two months after the intervention. The students of teachers who are not praised in treated schools are assigned lower marks two months after the intervention, but they do not perform any worse on final exams. Compared to costly interventions where teachers receive financial incentives, the effects of public praise for praised teachers are remarkably large.

Keywords: Public praise; Non-monetary incentives; Field experiment; Teacher performance.

JEL codes: C93, I21, J3, J45, J53, M52.

[*] Corresponding author. Permanent address: Centre for Economic Performance, The London School of Economics and Political Science. E-mail: m.a.cotofan@lse.ac.uk;

1. Introduction

Non-monetary incentives are playing an increasingly important role in many firms (Gallus and Frey, 2016). Praise, in particular, now features extensively in popular publications and the business literature as an effective way to motivate employees (see e.g. Nelson (2012)). A growing body of experimental research provides evidence for a positive effect of praise on performance (Stajkovic and Luthans, 2003; Grant and Gino, 2010; Kosfeld and Neckermann, 2011; Anderson et al., 2013; Ashraf et al., 2014; Lourenço, 2015; Bradler et al., 2016; Gallus, 2016; Gubler et al., 2016; Hoogveld and Zubanov, 2017). However, the existing evidence is predominantly confined to short-run effects in jobs involving simple and repetitive tasks. In this paper, I contribute to this body of literature by designing a large-scale field experiment to investigate the effects of public praise on performance. I study this question in a setting where employees - 900 teachers in 39 Romanian schools - perform cognitively complex tasks.

There is a growing literature on the effects of providing teacher incentives aimed at improving educational outcomes. However, empirical papers have focused almost exclusively on monetary incentives (for an overview, see Neal (2011)) and have found mixed effects on student performance (Leigh, 2012). Some studies have found positive effects of teacher incentives on student test scores and teacher attendance (Lavy, 2002, 2009; Duflo and Hanna, 2005; Glewwe et al., 2010; Muralidharan and Sundararaman, 2011; Imberman and Lovenheim, 2015). However, Springer et al. (2011) and Fryer (2013) study large-scale and costly interventions in the US, and find no treatment effects. While providing monetary incentives can increase teacher effort and can lead to better student performance, it can also crowd out teacher intrinsic motivation (Firestone and Pennell, 1993). What's more, if the incentive scheme is too complex and teachers feel as if they have little control, interventions may have no impact on student achievement (Fryer, 2013).

Little is known, however, about how effective non-monetary incentives such as public praise can be to improve teacher performance. A common concern with providing performance-based rewards is that once incentives are conditioned on performance, the performance measure becomes unreliable if it can be manipulated (Holmstrom and Milgrom, 1991). One solution is to link incentives to objective performance measures that cannot be manipulated. However, objective performance measures might not be frequently accessible or they may be entirely unavailable. Consequently, it is important to understand if non-monetary rewards also lead to gaming the performance measure on which they are conditioned or if lower-powered incentives are less susceptible to this. In this paper, I exploit a combination of subjective and objective performance measures to assess the effect of public praise for teachers.

I setup a randomized intervention in which teachers in treated schools are publicly praised based on improvements in the performance of their students. Prior to the intervention, these teacher assessments are a good predictor of student performance on anonymously marked standardized exams. While changes in teacher-assigned student grades can be easily calculated and frequently measured, they can also be manipulated by teachers in response to the intervention. In a sample of 900 teachers in 39 Romanian schools, I rank teachers based on changes in the grades of their students, within their own subject and across all schools. The 25% best teachers are labeled as top performers and qualify for praise. I exploit the fact that all schools in the sample use an online platform environment to publicly praise the top performing teachers in a random half of these schools. In the other half, no praise is provided. In treated schools, the intervention gives teachers a very coarse partition of their rank, namely whether they are in the top 25%, or not. In control schools, teachers do not receive any information.

The intervention is repeated twice more in the treated schools, at regular time intervals, throughout the remainder of the academic year. Empirical evidence suggests that announced praise increases the performance of all individuals (Kosfeld and Neckermann, 2011), while unannounced praise has a positive ex-post effect on the performance of non-recipients (Bradler et al., 2016;

Hoogveld and Zubanov, 2017). How a combination of the two impacts behavior remains entirely unexplored. In this paper, the effects of repeated praise are largely exploratory: by the end of the school year all teachers in treated schools are exposed to a combination of unannounced, announced, and repeated praise. In the appendix I present a descriptive analysis of how individuals respond to being repeatedly praised and how effective the intervention is once teachers learn to expect it.

I measure the effect of public praise on an objective performance measure: namely, student performance on anonymously marked standardized exams, undertaken by final year students six months after the first intervention. While the average treatment effect of public praise is insignificant, unexpected public praise raises the exam results of praised teachers' students by 0.17 standard deviations. The persistence and magnitude of the effect is remarkable given that public praise is not linked to exam performance.

As a robustness check, I also consider the effect of the intervention on teacher-assigned grades. I find that teachers in treated schools do not appear to systematically manipulate the grading of their students after being exposed to public praise: these assessments remain equally predictive of exam performance in treated schools, after the intervention. At the school level, unannounced praise does not have a statistically significant effect on teacher assessments. However, the point estimate is negative and economically significant: in treated schools teacher assessments decrease by 0.15 standard deviations. This average treatment effect is driven by opposite responses from the recipients and the non-recipients of praise in treated schools. The grades assigned by praised teachers increase by 0.23 standard deviations as compared to similar teachers in the control group, an increase also reflected in student exam performance. On the other hand, the grades assigned by non-praised teachers in the treatment group decrease by 0.30 standard deviations as compared to similar teachers in the control group, a decrease which is not reflected in exam performance.

The remainder of this paper is organized as follows. Section 2 introduces the setting, Section 3 describes the experimental design, Section 4 presents the main results, and Section 5 concludes.

2. Setting

The experiment targets nearly 900 teachers in 39 Romanian schools, who in total teach over 19,000 students aged 11 to 18. In Romania, the education system runs through three 4-year, pre-university education cycles: primary school (aged 7-10), secondary school (aged 11-14), and high school (aged 15-18). This experiment focuses on teachers from secondary schools and high schools.

Romania has a centralized education system, and schools follow the academic curriculum designed by the Ministry of Education. The curriculum provides a detailed guideline of the teaching material. Schools use comparable textbooks that are approved by the Ministry of Education, ensuring that teachers use the same materials and run through the curriculum in a similar order. As such, schools are homogeneous with respect to the type of information that students learn, and the competencies and skills they are expected to acquire throughout the school year. This experiment focuses on teachers of one of the following academic subjects: Romanian language, English language, Mathematics, Physics, Chemistry, Biology, History, Geography, and Computer Science. There is no evidence that teachers in Romania grade on a curve and they are certainly not required to do so. As such, grades are generally meant to reflect objective student learning. Figure A3 in the Appendix plots the distribution of student baseline grades across all schools.

At age 14 and at age 18, students are required to undertake standardized national-level examinations. These standardized exams are high-stake, as they help determine high-school and university admission. Undertaken in strictly invigilated exam centres, students work under the supervision of exam inspectors and exams are marked through a double-blind procedure. Thus, class teachers cannot influence their students' performance on these tests by either designing the test, helping students during the examination, or by marking the exam.

Teachers' wages and promotions are independent of their students' performance. Typically, teachers are subjected to standardized examinations and procedures to earn the right to be hired (*examen de titularizare* for becoming a teacher) or promoted (*gradul didactic I/II*), which are not contingent on student performance. As a consequence of that, there is no career incentive for teachers to artificially inflate the grades they assign to students, since they cannot get fired and will not be promoted based on this measure. This unique setting allows for cleanly identifying the effect of non-monetary incentives, as teachers cannot leverage praise to gain future monetary benefits.

The Ministry of Education strongly encourages -and provides guidelines to - teachers to administer a test in the beginning of the school year which measures the baseline ability of students. As such, I use the first grade that students receive from their teachers in the beginning of the academic year as a proxy for their baseline ability. In section 3.2, I show that this appears to be a reliable measure.

3. Experimental Design

This experiment follows 39 schools, located in 15 different regions in Romania.[1] Table A.1 in the Appendix provides some descriptive statistics of schools, teachers, and students in this experiment. All the schools in this experiment make use of an online education platform which tracks student progress. The platform is privately owned and administered, and it is not linked to either government officials nor to school principals. Schools can decide for themselves whether they want to implement the system and the usage of the platform comes at a small monthly cost.

The platform makes it easier for parents to keep track of their childrens' performance and

[1] The 39 schools in this experiment perform slightly better than the national average (a recent report on national performance can be found at https://www.edu.ro/rapoarte-publice-periodice). However, this is not a threat to the internal validity of the experiment as all the schools using the platform are randomly assigned to either treatment or control, and no schools drop out of the experiment.

attendance, as they are regularly posted online by teachers. By working directly with the platform providers and not with individual schools, I avoid any selection effects. Consequently, the setting qualifies as a natural field experiment, following the terminology in Harrison and List (2004). Access to the anonymized data allows me to monitor the performance of all students and teachers in the school for an entire academic year.

Schools are randomly assigned to either the treatment or the control group. Schools that are assigned to the treatment group receive "public praise", given through a message posted on the platform of each of the treated schools. Each message is posted by the managers of the company which provides the online platform, and is only visible to members of that school. The role of the platform managers is to oversee and manage the data from schools, and they do not have any stake in the employment and assessment of teachers, nor in the performance of the schools.

The message publicly praises the "best performing teachers" in each treated school. The best performing teachers are those who score among the top 25% across all schools, within their own subject. The first intervention is unannounced and it states that public praise will be repeated in the future. However, the exact date and frequency of future interventions is not disclosed. Subsequent rounds of praise take place at regular time-intervals until the end of the academic year. Appendix A.2 provides further details on the experimental time-line.

3.1 Intervention

After a period of collecting data on teacher and student baseline performance, the first intervention took place in January 2018. The messages were unannounced and unanticipated. In the schools that were assigned to the treatment group, a message (for the full intervention text, see Appendix A.3) is posted on the front page of each school's platform environment. The message is visible to all those who have a user account (teachers, parents, and students) immediately as they log-in.

The message is addressed to teachers and it states that the platform is interested in how student performance has improved since the beginning of the school year, as it is one of the ways to measure academic progress. The message announces that for a number of academic subjects platform managers have assessed the improvement in student grades across all the schools that use the electronic platform. They are told that based on this assessment, a number of teachers in their school are among the top 25% performers within their subject, across all the schools using the platform. The top performing teachers are listed by name, and thanked for their effort and contribution. Finally, the announcement mentions that such messages will be sent again in the future, to show the platform's gratitude towards teachers' hard work.

To further ensure that all teachers read the message carefully, an additional private message is sent to the personal inbox of all teachers in treated schools. The e-mail reminds teachers in treated schools about the intervention and provides them with a link to the public message (see Appendix A.3 for the full text).

The same procedure is repeated twice more throughout the remainder of the academic year, in March and in May of 2018. Following each intervention, teacher performance which is measured by changes in teacher-assigned student grades is computed on intervals of roughly two months.

3.2 Determining Top Performing Teachers

To determine the top performing teachers, I rank teachers based on changes in the grades of their students. While this approach is common in studies measuring the effectiveness of teacher incentives (Barrera-Osorio and Raju, 2017; Behrman et al., 2015; Glewwe et al., 2010; Muralidharan and Sundararaman, 2011), in this setting teachers have some freedom in assigning the grades. I further discuss this in section 3.3.

The school year is divided into four periods. Teacher performance is computed for each one of these time periods, namely before each of the three rounds of public praise, and once after the third and final round. Teachers are ranked according to an average of all the individual changes in the grades of their students (G_{it}). Each period, the change in the grades of student i is the difference between their baseline grades that period (denoted by θ_{ibt}) and their grades at the end of that period (θ_{it}):

$$G_{it} = \theta_{it} - \theta_{i_b t}.$$

θ_{it} is a weighted average of all the subsequent grades of a student within each period, where the final grade is given a weight of 50% and for all other intermediate grades, the remaining weight is equally distributed. The final grade is given a higher weight because it measures the longest period of time to pass since the baseline performance grade was recorded.[2]

Each round, the new baseline performance is replaced by the performance in the previous period, such that:[3]

$$\theta_{i_b t+1} = \theta_{it} = G_{it} + \theta_{i_b t}$$

The baseline performance for the first period ($\theta_{i_b t=1}$) is given by the first grade at the beginning of the school year and is a proxy for student ability. In Table A.4 in the Appendix, I show that well performing teachers do not increase the baseline performance of their students across school years and, as a consequence, are not mechanically less likely to be labelled as top performers in my experiment.

Table A.5 in the Appendix shows that a one standard deviation increase in pre-intervention

[2] This weighting method was requested by the educational experts managing the online platform during the design phase of the experiment and it has been pre-registered in the experimental design. Using a simple weighted average instead results into 94% of teachers being allocated on the same side of the threshold. As the exact formula is not communicated to teachers, this weighting decision should have no impact on the treatment effects.

[3] When θ_{it} is missing, θ_{it-1} will be used, and so on. If no previous average exists, $\theta_{i_b,t=1}$ is used.

teacher-assigned grades translates into a 0.32 standard deviations increase in performance on standardized exams, suggesting that prior to the intervention, changes in teacher-assigned grades are a good predictor of student performance on standardized exams.

Table 1 presents the average changes in teacher-assigned grades per academic subject, across all schools. For ease of interpretation, the measure is standardized, with a mean of zero and a standard deviation of one.

[Table 1 about here]

A teachers' performance based on this measure is an average of all the individual grade changes of their students, in the period in question. A teacher qualifies for public praise if, based on their students' grade changes, they are ranked in the top 25% best performing teachers, within their own subject. Top performing teachers at schools assigned to the treatment group are publicly praised. There are no treated schools in which no teacher is publicly praised, at any point throughout the experiment. The variation in teacher quality across schools is fairly comparable, with a standard deviation of 0.13.

3.3 Assessing the effects of public praise

The effect of public praise on teacher performance is measured by student performance on standardized exams. These exams take place at the end of the school year, for a subset of students ending an academic cycle, aged 14 and 18. Exams are marked by teachers from a different school through a double-blind procedure, such that a student's teacher cannot design the tests nor influence the grading. Since performance on standardized exams is not linked to public praise, teachers do not have a direct incentive to focus on this measure. If teachers who are praised increase effort, this should be reflected in their students' exam performance.

As a robustness check, I also analyze changes in teacher-assigned grades. This measure is directly linked to public praise: teachers qualify for praise based on these grades and they are told that future rounds of praise will also be conditioned on them. As such, it is important to understand the extent to which teachers manipulate the performance measure linked to public praise in order to be able to fully assess the effectiveness of such soft incentives. This is particularly relevant when truly objective performance measures are not available, which is frequently the case in the field. This combination of objective and subjective performance measures can help understand the trade-offs associated with non-monetary incentives.

3.4 Data and Randomization

The data spans 39 schools from 15 Romanian counties. Data collection records the performance of all the students in the school, across the 9 academic subjects of interest. In total, there are 855 teachers[4] in the sample and 19,748 students. Since each student takes on average roughly 7 of the 9 academic subjects,[5] there are in total 130,316 data entries.

Randomization is performed at the level of the treated unit: namely, the school, and stratified across three dimensions:

(i) Student baseline performance : A school-level weighted average of the initial grade that students receive at the beginning of the school year across all subjects, and a proxy for the average student ability in the school.

(ii) Teacher baseline performance: A school-level weighted average of the pre-intervention changes in teacher-assigned grades, and a proxy for the average teacher quality in the school.

[4] 13% of the teachers never record any grades, indicating some selection on the "type of teacher" that uses the platform. However, these teachers are similarly distributed between treated and control schools (p-value= 0.455), ensuring the internal validity of the experiment.

[5] Some subjects are only introduced in later years, and some students only choose, for example, a subset of science subjects.

(iii) School size: The number of teachers in the school (who actively use the platform and teach academic subjects).

Together, these three stratification variables capture the main sources of heterogeneity across the 39 schools, which are otherwise highly similar. Due to the limited number of schools, stratification variables are re-coded as binary indicators (above and below the sample average), as opposed to continuous measures. Within each strata, I randomly assign the 39 schools to either treatment or control. Due to a strata with just one school and the splitting of ties in favor of the treatment group, the randomization process assigned 21 schools (55% of teachers in the sample) to the treatment group and 18 schools (45% of teachers in the sample) to the control group. In all the main regression tables I control for the stratification variables and present p-values obtained through randomization inference which take the treatment assignment probabilities of each strata into account. In Table A.15 in the Appendix I show that the main results are also robust to directly including the treatment assignment propensity score as a control variable. [6]

Table 2 compares schools assigned to the treatment group with schools assigned to the control group. There are no significant differences in terms of either the stratification variables[7] or a number of additional important controls. Table A.7 in the Appendix shows that performing the balance tests at the teacher level provides similar results.

[Table 2 about here]

For the average teacher, changes in teacher-assigned grades are calculated based on 140 out of their 230 students, because not all the students in the sample have already been awarded multiple grades prior to the intervention.[8] There is no evidence that teachers in treated schools start testing

[6] As the randomization procedure assigned a propensity score of 1 to one school, this single-unit strata is dropped in this robustness check.

[7] To capture potentially fine grained differences, the continuous stratification variables are used in Table 2, as opposed to the binary indicators.

[8] The p-value on the difference between treated and control schools is 0.782.

their students more frequently post-intervention.[9] Furthermore, since switching class virtually never happens during the academic school year, selection effects due to students targeting publicly praised teachers are not a concern in this setting.[10]

There are 855 active teachers in the pre-intervention sample for whom changes in teacher-assigned grades are calculated. For 821 (96%) of them this measure is also calculated in the second round of intervention, for 758 (89%) of them in the third one, and for 729 (85%) of them in the last round. This attrition is not due to teachers leaving the school, but because a small number of teachers did not assess any of their students in the two months between interventions. Appendix A.6 shows that this attrition does not depend on being assigned to the treatment, on whether a teacher was a top performer or not, nor on the interaction between the two. Throughout the entire academic year, a total of 467 teachers are labelled as top performers at different points in time and a total of 248 teachers in treated schools receive public praise at least once.

4. Results of unannounced public praise

4.1 Standardized exams

To assess the effects of unannounced praise, I make use of results on the standardized exams that final year students undertake at the end of the school year. I estimate the following equation:

$$Exam_{ij} = \beta_1 T_{ij} + \beta_2 Top_j + \beta_3 T_{ij} * Top_j + \beta_4 X_{ij} + \nu_{ij} \tag{1}$$

where $Exam_{ij}$ is the final exam performance of student i, who has teacher j for that specific exam

[9] Calculated by looking at the difference in the number of recorded new grades per student after the intervention. The p-value for the coefficient that regresses the number of new grades after the first round on the treatment dummy is 0.686.

[10] Schools typically have a policy of not allowing students to sort into a different class than the one they were originally assigned to.

subject.[11] T_{ij} is the treatment dummy and Top_j is an indicator for whether that student's teacher was a top performer prior to the first intervention. The interaction term $T_{ij} * Top_j$ captures the effect of having a teacher who was praised in the first round. X_{ij} is a vector of controls at the student-level (gender, year of study, baseline student performance for the subject, track), teacher-level (subject), and school-level (region, whether in a rural area, whether publicly funded, size, baseline student and teacher performance at the school level, and past exam performance of the school), while ν_{ij} is an error term.

The standard errors are clustered at the school level. Due to the limited number of schools, for all the main results I report two types of p-values on the estimated coefficients. First, I report p-values calculated using the wild bootstrap procedure suggested by Cameron et al. (2008) with 1000 replications and clustering standard errors at the school level, which estimates reliable standard errors even with a small number of clusters. Second, I report permutation-based p-values which capture the probability of observing the estimated treatment effects, assuming that praise did not have an impact on teachers. In other words, these p-values compare the observed estimated treatment effects to their permutation distribution and are computed by repeatedly re-doing the random assignment (including the stratification) while clustering standard errors at the school level, with 1000 replications. The latter robustness test deals with uncertainty in estimates arising naturally from the random assignment of the treatments, while the former approach addresses uncertainty over the specific sampling from a large population (Athey and Imbens, 2017).

In total, 3,423 students are matched to the their exam marks, equivalent to 75% of the total number of final year students in the sample.[12] The students belong to 335 teachers from the original sample who teach final year classes. To asses whether the results can be generalized to all

[11] Students take two exams at the age of 14, in Mathematics and Romanian Language. At the age of 18, students take three exams: one in Romanian language, one in a compulsory track-specific subject, and one in a track-specific subject of their choice.

[12] The matching success rate is contingent on the name of the student being spelled identically in both the school-level database, and the database containing the exam marks which is provided by the Ministry of Education. The fully random matching errors do not impact the composition of the sample.

students, I show that final year students do not differ from other students on observables, neither are they more likely to be over-sampled from the treated group (see Appendix Table A.8 for details). Consequently, despite only having exam data for final year students, the results are likely to generalize to the whole sample.

In Table 3 below I assess the effect of public praise on the exam performance of students.[13] Figure A.4 in the Appendix plots the distribution of exam grades in treated and control schools. Table 3 presents the main results from estimating equation (1) and Table A.10 in the Appendix shows the estimated coefficients for all the control variables.

The β_1 coefficient in Column 1 of Table 3 captures the average treatment effect, and shows that the intervention has no statistically significant effect at the school level. In Column 2, the Treatment variable is interacted with "Top Performer" status. There are three parameters of interest. First, the treatment effect on the students of non-praised teachers is given by β_1, which is equivalent to the gap between bottom performing teachers in treated and control schools. β_1 also captures the spillover effect of public praise on non-praised teachers. Second, the treatment effect on the students of praised teachers is given by $\beta_1 + \beta_3$, which is equivalent to the gap between top performing teachers in treated and control schools. Finally, the difference in treatment effects between students of praised and non-praised teachers is given by β_3 and is equivalent to the gap between praised and non-praised teachers in treated schools, minus the corresponding gap in control schools. β_3 also captures the extent to which the intervention increased the gap between high- and low-performing teachers.

In Table 3, β_1 is equal to -0.089, and is not statistically different from 0 at any conventional level. Both bootstrapped (in brackets) and permuted (in braces) p-values confirm this. This shows that when it comes to student performance on standardized exams, there are little to no spillovers

[13] Table A.9 in the Appendix shows that the results are qualitatively the same if the analysis is performed through unweighted regressions at the teacher level.

from public praise on non-praised teachers in treated schools.[14] The effect of public praise on the praised, captured by $\beta_1 + \beta_3$, shows that students whose teacher was praised in the first round score 0.17 standard deviations higher on their final exams. The permuted p-value on the null hypothesis that $\beta_1 + \beta_3 = 0$ is 0.094. Finally, $\beta_3 = 0.258$ and indicates that public praise increases the gap between high- and low-performing teachers in treated schools — a coefficient that is statistically significant based on both bootstrapped and permuted p-values.

[Table 3 about here]

In a meta-analysis of studies on interventions in education, Sanders et al. (2015) find that effect sizes are usually no larger than 0.17 standard deviations. Thus, the magnitude of the effects on student achievement following a simple and cheap non-monetary incentive scheme for teachers are remarkable. This effect is especially large (and consistent with additional effort on the side of praised teachers) given the relatively small β_2 coefficient in Table 3, which indicates that in control schools, the value added of top performers is relatively small in the period following the intervention.

The empirical papers that isolate a causal effect of offering teachers monetary incentives provide an interesting comparison. Duflo and Hanna (2005) find an increase of 0.17 standard deviations in student achievement when teacher's pay is conditioned on their attendance. Muralidharan and Sundararaman (2011) find that incentive pay increased student achievement by 0.17 standard deviations in language and 0.27 standard deviations in math, during the first year. Glewwe et al. (2010) condition teacher incentive pay on the test scores of students and find that although the intervention increases student performance initially, the effects disappear once the program is discontinued. Finally, both Springer et al. (2011) and Fryer (2013) show that offering teachers large

[14] The small and insignificant β_2 coefficient suggests that self-assessed grades could be an imperfect measure of teacher performance. This may partly explain the negative point estimate for β_1 and it indicates that praise itself is a powerful motivator for top performers, rather than the informational aspect of the intervention.

transfers as a reward for improving student performance has no effect.

The effect of unannounced public praise on the exam performance of students of praised teachers is similar in magnitude to the effects of incentive pay found by Duflo and Hanna (2005) and Muralidharan and Sundararaman (2011), and much larger than the estimates of Springer et al. (2011) and Fryer (2013). Most importantly, public praise leads to better performance on standardized exams even though the reward itself is not conditioned on how well students do on these tests. This is in contrast to the findings of Glewwe et al. (2010) who argue that teachers only focused on the measures used to determine the rewards and put little effort in improving scores on exams not linked to incentives. This suggests that low-powered non-monetary incentives might lead to less gaming and be more likely to result in teachers having a broader focus on improving student outcomes. However, the positive effects of public praise in this setting are limited to praised teachers, while the average treatment effect remains insignificant. As such, one-to-one comparisons with previous studies measuring the overall impact of incentives are less clear-cut.

The evidence on the long-run effects of teacher incentives is almost entirely non-existent. In a notable exception, Lavy (2020) shows that monetary incentives for teachers lead to long-term gains in educational attainment, employment, and earnings. Evidence of real learning on the side of the students as a result of higher teacher effort suggests potential positive long-run consequences to soft-incentives, a promising avenue for future studies to explore.

Appendix A.11 undertakes an exploratory analysis on the relationship between repeated public praise and student exam performance. However, due to the endogenous nature of repeated public praise, the discussion remains largely speculative.

4.2 Changes in teacher-assigned grades

While student performance on standardized exams is a clean measure of teacher performance, public praise can incentivize teachers to manipulate the way they assign grades, after the intervention. An extensive cross-disciplinary literature argues that once a measure of performance is used to reward or to monitor, it becomes less effective over time (Goodhart, 1984). In other words, since in treated schools public praise is conditioned on teacher-assigned grades, the measure can become less informative about actual performance over time because teachers have the freedom to grade their own students.

To test this hypothesis, I look at the relationship between changes in teacher-assigned grades and exams over time, by performing a difference in difference analysis, at the teacher level, on the correlation between the objective performance measure of each of their student *i* (the standardized exam mark given by $exam_{ij}$) and the subjective incentivized measure of performance of each of their student *i* (changes in teacher-assigned grades given by g_{it}). I estimate the following equation:

$$Corr(exam_{ij}, g_{it}) = \lambda_0 + \lambda_1 Post_{it} + \lambda_2 T_{it} + \lambda_3 Post_{it} * T_{it} + \lambda_4 X_{it} + \varepsilon_{it}$$

where the correlation is calculated at the teacher level, across all of their students who undertake the standardized exam. $Post_{it}$ is an indicator for the post-intervention period, T_{it} is the treatment dummy, and X_{it} is a vector of controls. This approach is similar to Lavy (2009) who looks for evidence of teacher manipulation when teachers are rewarded based on a combination of objective and subjective metrics. The coefficient in Table 4 shows that changes in teacher-assigned grades do not become less predictive of exam performance in the treated group, after the introduction of public praise. Although this setting differs across a number of dimensions, the results are similar to Lavy (2009) who finds no evidence of teacher manipulation in response to incentive pay.

[Table 4 about here]

While these results do not exclude the possibility that some score manipulation still occurs, I assess teachers' behavioural response to unannounced praise by analysing changes in the way they grade their students following the intervention. I estimate a before-after comparison using the following equation:

$$G_{it+1} = \alpha_1 T_{it} + \alpha_2 Top_{it} + \alpha_3 T_{it} * Top_{it} + \mu_i + \tau_t + \varepsilon_i \tag{2}$$

where G_{it+1} is the change in teacher assigned grades, calculated as outlined in section 3.2. T_{it} is a treatment dummy, Top_{it} is an indicator for being a top performer, and $T_{it} * Top_{it}$ is the interaction between being a top performer and being in a treated school. μ_i is a teacher-specific fixed effect which captures all time-invariant teacher characteristics and τ_t is a time fixed effect. The analysis is performed at the teacher level and the standard errors are clustered at the school level. In line with the approach detailed in section 4.1, more conservative bootstrapped p-values and permutation-based p-values are reported alongside all coefficients.

Figure A.5 in the Appendix plots the distribution of teacher-assigned grades in treated and control schools, prior and post intervention. Table 5 estimates equation (2).

[Table 5 about here]

Column 1 shows that at the school level the point-estimate of the average treatment effect is negative, with teachers in treated schools performing 0.15 standard deviations worse than similar teachers in the control group. Although the effect is not statistically significant at any conventional level, it is economically significant and policy relevant.

Column 2 decomposes the average treatment effect, by interacting it with "Top Performer" status, and shows that it is driven by opposing responses from top and bottom performing teachers. α_1 is the treatment effect on non-praised teachers in treated schools and captures the spillover effect of public praise on non-praised teachers. These teachers decrease performance by 0.30 standard deviations, as compared to bottom performing teachers in the control group, a statistically significant effect as confirmed by both bootstrapped and permuted p-values. $\alpha_1 + \alpha_3$ is the treatment effect on the praised teachers, who increase performance by 0.23 standard deviations as opposed to top-performing teachers in the control group. Using randomization inference with 1000 replications, the permuted p-value on the null hypothesis that $\alpha_1 + \alpha_3 = 0$ is 0.192. Finally, the difference in treatment effects between students of praised and non-praised teachers is given by α_3 and is equivalent to the gap between praised and non-praised teachers in treated schools, minus the corresponding gap in control schools. The α_3 coefficient is statistically significant based on both bootstrapped and permuted p-values and shows that in terms of teacher assigned grades, the intervention increased the gap between high- and low-performing teachers significantly.

The results show that unannounced praise has a significant effect on the teacher-assigned student grades. While it cannot be ruled-out that teachers are to some extent changing their assessments in response to public praise, at least part of the positive effect on the changes in teacher-assigned grades (0.23 standard deviations in Table 5) is also reflected in the exam performance of students (0.17 standard deviations in Table 3)[15] and persists six months after the intervention.

The negative effect of not being praised on changes in teacher-assigned grades (0.30 standard deviations in Table 5) does not reflect final exam performance (-0.089 standard deviations in Table 3) in a statistically significant way. Furthermore, in Table A.13 in the Appendix I split bottom performers into three quantiles and show that they all decrease performance to a similar extent, following the intervention. This confirms that the results are not driven by teachers ranked just below the threshold, who may be more susceptible to being disappointed by missing out on praise, but that all teachers in the bottom category respond in a similar fashion.

[15] While the results in Table 5 include the entire sample of students, in Table A.8 in the Appendix I show that final year students are not significantly different across demographics.

In Appendix A.12, I descriptively show that repeated interventions do not appear to move teacher performance. This also holds for those teachers who are praised for the first time in a repeated intervention and is consistent with the idea that once rewards are expected and internalized they tend to become less efficient.

Finally, in Appendix A.14 I explore how the intervention impacted student attendance throughout the academic year. While these results provide further direct evidence on the impact of the treatment, they are subject to two important limitations. First, these results are also to some extent indirect because student attendance is recorded by the class teacher. Second, variation in this outcome measure is limited: as attendance is mandatory, student attendance tends to be very high across the board, such that the average student skips 3.24 hours of school throughout the entire academic year, with a standard deviation of 6.05 hours.

Table A.14.1 shows how student attendance differs by treatment status and by teacher performance. While the findings in Table A.14.1 fall short of any conventional statistical significance level, the sign of the coefficients points to praised teachers marginally increasing the attendance of their students. Table A.14.2 splits the results by high and low performing students. Again, all coefficients fall short of statistical significance. However, the sign of the coefficients indicate that praised teachers appear to focus on increasing the attendance of low-performing students. The findings in these tables appear to be consistent with the results in sections 4.1 and 4.2, and in line with additional effort on the side of praised teachers. However, given the low power and limited variation in attendance, and the lack of statistical significance, these results should be interpreted cautiously.

5. Concluding remarks

The results of this experiment speak to a growing interest in using low-powered incentives to increase the performance of workers. It also contributes to the scarce evidence on how praise affects workers performing cognitively complex tasks. To the best of my knowledge, this is also the first study to assess the effectiveness of non-monetary incentives for teachers and to show that public praise can have large and persistent effects on their performance. This experiment shows that public praise is effective at improving the performance of praised teachers. A common concern with rewarding teachers based on performance is that incentives based on subjective assessments can lead to gaming. The findings of this paper show that the increase in teacher performance due to praise cannot be entirely ascribed to score manipulation: students of praised teachers have both higher teacher-assigned grades and better performance on anonymously marked standardized exams. This is particularly remarkable given that praise is not linked to exam grades. For non-praised teachers, the treatment effect on teacher-assigned grades is negative and sizable. However, this does not reflect the exam performance of their students in a statistically significant manner, indicating that the negative effect of public praise may be less persistent than the positive effect.

Overall, the average treatment effect at the school-level provides a cautionary tale for employers and policy makers. More work is needed to be able to fully weigh the costs of such interventions against their benefits. From a cost perspective, further studies are needed to understand why public praise has a negative effect on the non-praised when it comes to teacher-assigned grades, and why this negative effect is not reflected in performance on standardized exams. The design of this experiment cannot clearly disentangle why non-praised teachers appear to become discouraged by the intervention in the short-term. Despite being a good predictor of exam performance, teacher-assigned grades may still be an imperfect measure of teacher performance. If so, the negative effect could be caused by non-praised teachers reacting to what they perceive as a flawed reward system. More objective measures of performance could attenuate this response.

Nonetheless, crowding out of intrinsic motivation may occur even when performance measures are perceived as entirely objective (Crutzen et al., 2013). Striking the right balance between low-powered incentives that are sufficiently salient to improve the performance of recipients without crowding out the intrinsic motivation of non-recipients appears to be crucial. Subsequent randomized control trials could explore different performance thresholds, experiment with the coarseness of disclosed rankings, vary the salience of the intervention, or explore a combination of private and public rewards to further evaluate the costs associated with public praise. From a benefits perspective, the picture is clearer: the results of this experiment indicate that public praise can be very effective at increasing the performance of those who are rewarded. With monetary incentives being vastly more expensive and sometimes inefficient, non-monetary rewards are an avenue worth further exploring.

Acknowledgements

This paper reports the results from a field experiment for which the design was pre-registered at $https : //www.socialscienceregistry.org/trials/2604/history/32360$.

I am grateful to all those who provided valuable comments and feedback on early drafts, especially to Robert Dur, Josse Deelfgauw, Bauke Visser, Anne Boring, Jan Stoop, and the participants in many conferences and seminar presentations.

References

Anderson, Ashton, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec, "Steering user behavior with badges," in "Proceedings of the 22nd international conference on World Wide Web" ACM 2013, pp. 95–106.

Ashraf, Nava, Oriana Bandiera, and Scott S Lee, "Awards unbundled: Evidence from a natural field experiment," *Journal of Economic Behavior & Organization*, 2014, *100*, 44–63.

Athey, Susan and Guido W Imbens, "The econometrics of randomized experiments," in "Handbook of economic field experiments," Vol. 1, Elsevier, 2017, pp. 73–140.

Barrera-Osorio, Felipe and Dhushyanth Raju, "Teacher performance pay: Experimental evidence from Pakistan," *Journal of Public Economics*, 2017, *148*, 75–91.

Behrman, Jere R, Susan W Parker, Petra E Todd, and Kenneth I Wolpin, "Aligning learning incentives of students and teachers: Results from a social experiment in Mexican high schools," *Journal of Political Economy*, 2015, *123* (2), 325–364.

Bradler, Christiane, Robert Dur, Susanne Neckermann, and Arjan Non, "Employee recognition and performance: A field experiment," *Management Science*, 2016, *62* (11), 3085–3099.

Cameron, A Colin, Jonah B Gelbach, and Douglas L Miller, "Bootstrap-based improvements for inference with clustered errors," *The Review of Economics and Statistics*, 2008, *90* (3), 414–427.

Crutzen, Benoît SY, Otto H Swank, and Bauke Visser, "Confidence management: on interpersonal comparisons in teams," *Journal of Economics & Management Strategy*, 2013, *22* (4), 744–767.

Duflo, Esther and Rema Hanna, “Monitoring works: Getting teachers to come to school,” Technical Report, National Bureau of Economic Research 2005.

Firestone, William A and James R Pennell, “Teacher commitment, working conditions, and differential incentive policies,” *Review of educational research*, 1993, *63* (4), 489–525.

Fryer, Roland G, “Teacher incentives and student achievement: Evidence from New York City public schools,” *Journal of Labor Economics*, 2013, *31* (2), 373–407.

Gallus, Jana, “Fostering public good contributions with symbolic awards: A large-scale natural field experiment at Wikipedia,” *Management Science*, 2016, *63* (12), 3999–4015.

_ **and Bruno S Frey**, “Awards as non-monetary incentives,” in “Evidence-based HRM: a Global Forum for Empirical Scholarship,” Vol. 4 Emerald Group Publishing Limited 2016, pp. 81–91.

Glewwe, Paul, Nauman Ilias, and Michael Kremer, “Teacher incentives,” *American Economic Journal: Applied Economics*, 2010, *2* (3), 205–27.

Goodhart, Charles AE, “Problems of monetary management: the UK experience,” in “Monetary Theory and Practice,” Springer, 1984, pp. 91–121.

Grant, Adam M and Francesca Gino, “A little thanks goes a long way: Explaining why gratitude expressions motivate prosocial behavior.,” *Journal of personality and social psychology*, 2010, *98* (6), 946.

Gubler, Timothy, Ian Larkin, and Lamar Pierce, “Motivational spillovers from awards: Crowding out in a multitasking environment,” *Organization Science*, 2016, *27* (2), 286–303.

Harrison, Glenn W and John A List, “Field experiments,” *Journal of Economic literature*, 2004, *42* (4), 1009–1055.

Heß, Simon, “Randomization inference with Stata: A guide and software,” *The Stata Journal*, 2017, *17* (3), 630–651.

Holmstrom, Bengt and Paul Milgrom, “Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design,” *JL Econ. & Org.*, 1991, *7*, 24.

Hoogveld, Nicky and Nick Zubanov, “The power of (no) recognition: Experimental evidence from the university classroom,” *Journal of Behavioral and Experimental Economics*, 2017, *67*, 75–84.

Imberman, Scott A and Michael F Lovenheim, “Incentive strength and teacher productivity: Evidence from a group-based teacher incentive pay system,” *Review of Economics and Statistics*, 2015, *97* (2), 364–386.

Kosfeld, Michael and Susanne Neckermann, “Getting more work for nothing? Symbolic awards and worker performance,” *American Economic Journal: Microeconomics*, 2011, *3* (3), 86–99.

Lavy, Victor, “Evaluating the effect of teachers’ group performance incentives on pupil achievement,” *Journal of political Economy*, 2002, *110* (6), 1286–1317.

_ , “Performance pay and teachers’ effort, productivity, and grading ethics,” *American Economic Review*, 2009, *99* (5), 1979–2011.

_ , “Teachers’ Pay for Performance in the Long-Run: The Dynamic Pattern of Treatment Effects on Students’ Educational and Labour Market Outcomes in Adulthood,” *The Review of Economic Studies*, 2020, *87* (5), 2322–2355.

Leigh, Andrew, “Teacher pay and teacher aptitude,” *Economics of education review*, 2012, *31* (3), 41–53.

Lourenço, Sofia M, "Monetary incentives, feedback, and recognition—Complements or substitutes? Evidence from a field experiment in a retail services company," *The Accounting Review*, 2015, *91* (1), 279–297.

Muralidharan, Karthik and Venkatesh Sundararaman, "Teacher performance pay: Experimental evidence from India," *Journal of political Economy*, 2011, *119* (1), 39–77.

Neal, Derek, "The design of performance pay in education," in "Handbook of the Economics of Education," Vol. 4, Elsevier, 2011, pp. 495–550.

Nelson, Bob, *1501 ways to reward employees*, Workman Publishing, 2012.

Roodman, David, Morten Arregaard Nielsen, James G. MacKinnon, and Matthew D. Webb, "Fast and wild: Bootstrap inference in Stata using boottest," *The Stata Journal*, 2019, *19* (1), 4–60.

Sanders, Michael, Aisling Ni Chonaire et al., ""Powered to Detect Small Effect Sizes": You keep saying that. I do not think it means what you think it means.," Technical Report, Department of Economics, University of Bristol, UK 2015.

Springer, Matthew G, Dale Ballou, Laura Hamilton, Vi-Nhuan Le, JR Lockwood, Daniel F McCaffrey, Matthew Pepper, and Brian M Stecher, "Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching (POINT).," *Society for Research on Educational Effectiveness*, 2011.

Stajkovic, Alexander D and Fred Luthans, "Behavioral management and task performance in organizations: conceptual background, meta-analysis, and test of alternative models," *Personnel Psychology*, 2003, *56* (1), 155–194.

Webb, Matthew D, "Reworking wild bootstrap based inference for clustered errors," *Queen's Economics Department Working Paper, No. 1315*, 2013.

Appendix

A.1 Descriptive Statistics

[Table A.1 about here]

A.2 Data Collection and Time Line

This experiment follows nearly 900 teachers in 39 Romanian schools over the course of an entire academic year, from September 2017 to August 2018. All the 39 schools in the experiment use an online management platform designed to monitor student performance. Between the 1st of September 2017 and the 21st of January 2018, data on the baseline performance of teachers is collected, and used to compute changes in teacher-assigned student grades.

The treatment assignment of schools remains unchanged throughout the entire academic year. There is no selection into the treatment, and no schools opt out of receiving the public praise messages following the interventions.

The first intervention takes place on the 22nd of January 2018, after the end of the winter break. The platform managers post the public praise messages on the platform page of each of the treated schools. The messages posted in each school are identical in terms of content. The only source of variation in the messages is the names of the top performing teachers within each school. The message is only visible to teachers, parents, and students within that school. The message is posted on the main page of the platform, visible immediately after logging-in. An additional email is sent by the platform managers to all teachers in the school, reminding them to read the public message and providing them with a link to the original post.

The second intervention takes place on the 20th of March 2018, two months after the first round of messages is sent. Student grades between the 22nd of January 2018 and the 20th of March 2018 are used to calculate the new changes in teacher-assigned student grades for all teachers across all schools. Teachers are ranked again based on this new measure, and top performing teachers in treated schools are publicly praised in a new round of messages posted on the 20th of March 2018.

The third and final intervention takes place on the 15th of May, two months after the second round of messages. Analogous to previous interventions, a final round of public messages is posted in all of the treated schools. Finally, changes in teacher-assigned student grades are calculated again between the 15th of May 2018 and the end of the academic year.

In the week of 11th of June 2018, students finishing secondary school (aged 14) undertake high stake standardized exams, which are anonymously marked ('Examen de capacitate'). In the week of 25th of June 2018, students finishing high school (aged 18) undertake high-stake standardized exams which are anonymously marked ('Examen de Bacalaureat').

Figure A.2 presents a schematic overview of the experimental design and timeline.

[Figure A.2 about here]

A.3 Full text of intervention

The intervention text, original in Romanian language, is posted by the platform managers on the front page of the website, visible to all teachers, parents and students immediately after logging-in.

Unannounced public praise message:

”Dear Teachers,

We are interested in how the performance of this school's students is improving over time, since we want to encourage progress in education.

One way to measure the progress of students, is to see how much their grades improved since the beginning of the year. For a number of subjects (Mathematics, Romanian, English, Biology, Chemistry, Physics, History, Geography and Computer Science) we have looked at the improvement in student grades across all the schools that implement the *[platform name]* school management solution.

We are happy to announce that a number of teachers in your school are among the top 25% performers for their subject, across all the schools in our database. For these subjects, their student's grades have improved the most since the beginning of the semester, as compared to the grades of students from other schools! These teachers are:

[Teacher 1 name]

.....

[Teacher n name]

We would like to thank these teachers in particular for their contribution!

In the future we plan to send such messages more often, to show our gratitude towards your hard work!

Best,"

Announced and repeated public praise message:

"Dear Teachers,

As you know, in the past we have analyzed, for a number of subjects (Mathematics, Romanian, English, Biology, Chemistry, Physics, History, Geography, English and Computer Science), the improvement in student grades across all the schools that implement the *[platform name]* school management solution.

We have now repeated this analysis. We are happy to announce that a number of teachers in your school are among the top 25% performers for their subject, across all the schools in our database. For these subjects, their student's grades have improved the most over the last 2 months, as compared to the grades of students from other schools! These teachers are:

[Teacher 1 name]

.....

[Teacher n name]

We would like to thank these teachers in particular for their contribution!

In the future we plan to send such messages more often, to show our gratitude towards your hard work!

Regards,"

Additionally, each teacher in a treated school is sent a reminder about the public message, through a personal e-mail from the platform managers. This measure is implemented to ensure that the treatment is as visible as possible.

Following the first intervention, the private message sent to all teachers is:

"Hello,

We are pleased to announce that for a number of subjects we have reviewed the increase in student performance in schools which implement the *[platform name]* school management solution.

Based on this analysis, *[number teachers]* teachers in your school are among the top 25% teachers in existing schools in our database!

If you want to see who these teachers are (or if you are one of them) you can see the list here: *[link to public message]*

Regards,"

Following the second and the third intervention, the private message sent to all teachers is:

"Hello,

We are pleased to announce that we have reviewed again the increase in student performance over the past 2 months. This analysis included all the schools which implement the *[platform name]* school management solution.

Based on this analysis, *[number teachers]* teachers in your school are among the top 25% teachers in existing schools in our database!

If you want to see who these teachers are (or if you are one of them) you can see the list here: *[link to public message]*

Regards,”

A.4 Changes in teacher-assigned grades: robustness checks

[Table A.4 about here]

A.5 Relationship between pre-intervention changes in teacher-assigned grades and exam performance

[Table A.5 about here]

A.6 Attrition

Attrition occurs when a teacher did not record any grades, for any of their students, during one experimental period. As such, for this teacher, changes in teacher-assigned student grades cannot be calculated. From the 855 teachers in the original sample, for 4% (n = 34) this measure cannot be calculated in period 2, 7% (n = 63) in period 3, and 3% (n = 29) in the last period. For the remaining teachers, changes in teacher-assigned student grades can be calculated throughout the intervention, making the attrition rate reasonably low.

Table A.6 below shows the results from the Hotelling's T-squared test for multivariate data. The test verifies whether two sets of means are equal to each other across two groups, namely between a group of teachers who opt-out by recording no grades ('Attrition') and a group of teachers who do not opt-out ('No attrition'). The test has the advantage of jointly testing multiple variables at the same time, in this case the treatment status, being a top performer, and the interaction between the two. In the case of only one variable, the test reduces to a standard t-test. According to the results in Table A.6, attrition each round does not depend on either treatment status, on being a top performer, or on the interaction between the two.

[Table A.6 about here]

A.7 Balance tests at the teacher level

Since the treatment assignment is at the level of the school, Table 2 in the main text presents the balance tests at the school level. However, given the relatively small number of schools, the number of observations is also limited. I additionally provide balance tables at the teacher level as well.

While some of the balancing variables are by definition determined at the school level (school size, the percentage of schools in urban areas, and the percentage of publicly funded schools), Table A.7 shows balancing tests also for variables calculated at the teacher level (baseline performance of a teacher's students, a teacher's baseline performance, the share of female students in a teacher's class, and the total number of skipped classes across all the students of a teacher). All the results are consistent with those in Table 2, as expected if random assignment is successful.

[Table A.7 about here]

A.8 Balance tests for final years students

Table A.8 reports the coefficients, standard errors and p-values from regressing a number of controls on a dummy variable which takes value one if a student is in the final year, and zero otherwise. With one exception, final year students do not appear to be different across any dimension, neither are they more likely to be over-sampled from the treated group. Final year students are slightly less likely to be sampled from schools that have some private funding. However, this is mechanically determined by the fact that these schools are mostly focused on secondary education, and typically do not offer classes for final year high school students. The main specifications control for school funding.

[Table A.8 about here]

A.9 The effect of unannounced public praise on standardized exam performance - teacher-level regressions

In this robustness check, the analysis in Table 3 is performed through unweighted regressions at the teacher level. The first column uses the full teacher sample. Since the the sample used to calculate the results in Table 3 is significantly smaller than the sample used to calculate the results in Table 5, there is a larger risk of over-weighting teachers with few students. Consequently, Columns 2 and 3 restrict the sample to avoid over-weighting teachers who have a small number of students taking the final exam. The second column drops teachers who teach a small number of students (below the median). The third column only keeps teachers of compulsory exams subjects (Mathematics and Romanian Language) which have a much larger number of students taking the final exam than teachers of elective subjects.

[Table A.9 about here]

A.10 The full set of coefficients for the effect of unannounced public praise on standardized exam performance

[Table A.10 about here]

A.11 Announced and repeated public praise: standardized exams

To explore the relationship between repeated public praise and final exam performance, I estimate:

$$Exam_{ij} = \phi_1 T_{ij} + \phi_2 Freq_j + \phi_3 T_{ij} * Freq_j + \phi_4 X_{ij} + \nu_{ij} \tag{3}$$

where $Exam_{ij}$ is the final exam performance of student i, under teacher j. $Freq_j$ is a categorical dummy which indicates whether, and how many times, a teacher was a top performer throughout the academic year. In other words, $Freq_j$ takes value 0 if a teacher was always a bottom performer, value 1 if a teacher was a top performer only once, and value 2 if a teacher was a top performer repeatedly throughout the year. The vector of controls X_{ij} is defined as in equation (1), and ν_{ij} is an error term. Standard errors are clustered at the school level.

In Table A.11 I estimate equation (3) which classifies teachers as always bottom performers (135 teachers and 2,192 students), top performers only once throughout the experiment (138 teachers and 2,281 students) and top performers more than once (45 teachers and 1,027 students).

Dis-aggregating the results by the frequency with which a teacher was a top performer reveals some heterogeneity. The students of those teachers who were never praised do not perform any different than their peers in the control group, with a point estimate of zero. The students of teachers who were only praised once throughout the academic year do not perform significantly different either. On the other hand, the students of repeatedly praised teachers perform 0.33 standard deviations better on final exams, as compared to their counterparts in the control group. The p-value from a simple t-test on the difference between ϕ_1 and $\phi_1 + \phi_{3,2}$ is smaller than 0.001. This finding appears to be predominantly driven by teachers who were praised in the first and in the third round (roughly 60% of the teachers who were praised repeatedly). In line with coefficient ψ_{10} in Table A.12.4, these findings suggest that in the long-run, repeated public praise can be an effective tool if given sparingly. However, those teachers who are top performers multiple times in the treated group are a select type, and have predominantly also been praised in the first round. Thus, the results should be interpreted accordingly and with caution.

[Table A.11 about here]

A.12 Announced and repeated public praise: changes in teacher-assigned student grades

To estimate the per-period coefficient on the combination of unannounced and announced praise on changes in teacher-assigned grades I use the following equation:

$$G_{it+1} = \sum_{t=1}^{t} \gamma_{1,t} * T_{it} + \sum_{i,t=1}^{t} \gamma_{2,t} * Top_{it} + \sum_{i,t=1}^{t} \gamma_{3,t} * T_{it} * Top_{it} + \mu_i + \tau_t + \omega_{it} \qquad (4)$$

where $t = 1$ is the pre-intervention period, and at $t = \{2, 3, 4\}$ the three intervention rounds take place. As such, at $t = 2$ top performers in treated schools receive unannounced praise. At $t = \{3, 4\}$ top performers in treated schools receive announced praise, and a subset of them receive repeated praise. Standard errors are clustered at the school level.

To explore how repeated public praise relates to teacher performance, I estimate (i) the coefficient on not being praised in a given round, (ii) the coefficient on being praised for the first time in any given round, and (iii) the coefficient on being praised repeatedly in any given round, as compared to similar teachers in the control group:

$$G_{it+1} = \delta_1 T_{it} + \sum_{i,j=0}^{2} \delta_{2,j} Type_{ijt} + \sum_{i,j=0}^{2} \delta_{3,j} T_{it} * Type_{ijt} + \mu_i + \tau_t + \psi_i \qquad (5)$$

where $Type_{ijt}$ is a categorical variable which records the type j of a teacher i within each period t. Specifically, $Type_{ijt}$ takes value 0 if teacher i is not a top performer at time t, value 1 if teacher i is a top performer for the first time at time t, and value 2 if teacher i is a top performer for the second or third time at time t.

Announced (and for some teachers repeated) praise is given two times throughout the remainder of the school year, namely two months and four months after unannounced public praise. After the second round, changes in teacher-assigned grades are calculated again for 89% of the active teachers in the original sample, and after the final round the measure is calculated for 86% of the active teachers in the original sample. As described in Appendix A.6, this attrition is random and does not relate to being in the treated group, to being a top performer, or to the interaction between the two.

The remainder of this section discusses how announced and repeated praise relates to changes in teacher-assigned grades. There are no differences between treatment and control in the number of new grades (per student) that teachers record, confirming that there is no gaming on the side of the teachers at the extensive margin.[16] To shed more light on the way learning is distributed throughout the academic year, Table A.12.1 shows the average changes in teacher-assigned grades across treated and control schools, throughout the experiment.

[Table A.12.1 about here]

Table A.12.2 provides an overview of the results across all periods, by estimating equation (4).

[16] Calculated by looking at the difference in the number of recorded new grades per number of students that a teacher has, after each round. The p-value for the coefficient that regresses the number of new grades on the treatment dummy is 0.125 after the second round and 0.947 after the third one.

Announced praise does not seem to have any relation to the performance of either top or bottom performers, a finding consistent in both repeated interventions. This is in line with the idea that once rewards are anticipated, they tend to lose their effectiveness in moving performance.

[Table A.12.2 about here]

To shed light on the relationship between repeated praise and teacher performance, I estimate equation (5) which compares the performance of teachers who were not praised in any given round, with the performance of teachers who were praised for the first time, and the performance of teachers who were praised for a repeated time within that round. Table A.12.3 presents the results. Praising teachers for the first time (in any round) does not translate into any changes in teacher-assigned grades in the following period. Since public praise in the first round has a large and significant effect on the performance of both top and bottom performing teachers, the small and insignificant coefficients δ_1 and $\delta_{3,1}$ suggests that the positive effects of public praise observed in Table 5 disappear when teachers anticipate the intervention. Being praised repeatedly does not appear to influence teacher performance and the point-estimate is negative.

[Table A.12.3 about here]

Equation (5) imposes the restrictive assumption that in treated schools the response of any teacher *i* at time *t* is independent of their experiences in previous rounds. However, repeating the intervention in treated schools and exposing the same group of teachers to multiple treatments gives rise to increasingly complex combinations of effects with each additional round. To relax this assumption, in Table A.12.4 I estimate a flexible specification controlling for each type of experience that a teacher could have had throughout the year, such that at each point in time the previous performance of a teacher is taken into account. The reference category is made up of teachers who were never praised, up to that period.

[Table A.12.4 about here]

While the effects of unannounced praise remain sizable and precisely estimated, repeated interventions do not appear to significantly affect teacher performance. However, some interesting patterns arise. First, teachers who are praised for the first time in subsequent rounds (ψ_4 and ψ_9) do not appear to improve performance. Second, those teachers who were only praised in the first round and became more motivated as a result of that, appear to exert additional effort to maintain a high performance throughout (ψ_3 and ψ_6). This is particularly visible following the final intervention, with a marginally significant increase in teacher-assigned grades, as compared to similar teachers in the control group. Third, there are no clear benefits from praising a teacher in two consecutive interventions (ψ_5, ψ_{11} and ψ_{12}).

These results suggest a number of additional takeaways. When rewards are given repeatedly, being 'first to the prize' seems to matter more. Teachers who are praised in the first round increase performance (ψ_2) and appear to remain more intrinsically motivated throughout the remainder of the experiment (ψ_3 and ψ_6). On the other hand, being second or third to the prize does not translate into better performance. Finally, repeated rewards over short periods of time do not achieve the desired results, as coefficients on being praised two or three periods consecutively are always negative (ψ_5, ψ_{11} and ψ_{12}). However, being praised in the first round and in the third one returns a large and positive coefficient (ψ_{10}) following the final intervention. This indicates that while praise looses bite as it becomes less scarce, it remains a powerful tool for those who receive it sparingly.

A.13 Heterogeneous treatment effects for bottom performers

I estimate the following equation:

$$G_{it+1} = \alpha_0 + \alpha_1 T_{it} + \alpha_2 Quant_{it} + \alpha_3 T_t * Quant_{it} + \mu_i + \tau_t + \varepsilon_{it} \qquad (6)$$

where G_{it+1} measures changes in teacher-assigned student grades following the first intervention, and $Quant_{it}$ is a set of dummies for each of the four quantiles of the teachers's performance distribution. Table A.13 below shows the results from estimating equation (6) for teachers in the 1st quantile who qualified for praise (the top 25%), teachers in the 2nd quantile (between 25% and 50% of the performance distribution), teachers in the 3rd quantile (between 50% and 75% of the performance distribution), and teachers in the 4th quantile (the bottom 75% of the distribution), each compared to similar teachers in the control group.

Table A.13 shows that teachers in treated schools at different quantiles of the bottom 75% performance distribution do not respond differently following the intervention. A test of joint equality of the coefficients on the three bottom quantiles returns an F-value of 1.90 and a p-value of 0.16.

[Table A.13 about here]

A.14 Student Attendance

Table A.14.1 shows how student attendance differs by treatment status and by teacher performance. In column (1), the outcome variable is the change in the number of skipped classes, two months after the intervention. In column (2) the outcome variable is the change in the number of skipped classes, by the end of the school year. While the findings in Table A.14.1 fall short of any conventional statistical significance level, the sign of the coefficients points to praised teachers marginally increasing the attendance of their students. By the end of the school year, the students of both praised and non-praised teachers in treated schools appear to attend class slightly more frequently. In column (2), the permuted p-value on Alpha1+Alpha3 is equal to 0.302.

[Table A.14.1 about here]

To further explore this mechanism, Table A.14.2 splits the results by high and low performing students. Again, all coefficients fall short of statistical significance. However, the sign of the coefficients indicate that praised teachers appear to focus on increasing the attendance of low-performing students (Alpha3 in columns 1 and 2). The magnitude of the coefficient is also sizable, at 0.178 standard deviations by the end of the school year. In column 2, the permuted p-value on Alpha1+Alpha3 is equal to 0.264.

[Table A.14.2 about here]

A.15 Effects on exam performance with controls for the propensity score

[Table A.15 about here]

Tables and Figures

Table 1: Average changes in teacher-assigned student grades per academic subject, in the beginning of the school year

Subject	**Mean**	**Standard Deviation**	**No. teachers**
Biology	0.151	0.898	64
Chemistry	0.008	1.398	46
Computer Science	-0.050	1.036	60
English Language	-0.077	0.894	145
Geography	0.143	0.880	66
History	-0.016	0.955	64
Mathematics	-0.045	0.948	151
Physics	0.132	1.334	85
Romanian Language	-0.050	0.896	174

Notes: Columns show the mean and the standard deviation of changes in teacher-assigned grades across all subjects, prior to the intervention. The measure is standardized, with a mean of zero and a standard deviation of one.

Table 2: Balance tests for mean differences between treatment and control

Variable	**C**	**T**	**P-value**
Student baseline performance	-0.073	0.062	0.681
	(0.234)	(0.224)	
Teacher baseline performance	-0.127	0.109	0.472
	(0.223)	(0.230)	
School size (no. teachers)	21.611	22.238	0.890
	(3.240)	(3.130)	
% Urban schools	0.833	0.810	0.852
	(0.090)	(0.088)	
% Publicly funded	0.833	0.762	0.594
	(0.090)	(0.095)	
% Female students	0.524	0.542	0.632
	(0.027)	(0.023)	
No. skipped classes	0.745	0.650	0.585
	(0.139)	(0.105)	
N	18	21	

*Notes: The balance tests are performed at the school level. The first two columns show variable means between the control group of schools, and the treated group of schools. In brackets, standard deviations are presented. The third column shows the p-values from two-sample t-tests on the null hypothesis that group means are equal. Student baseline performance is standardized with a mean of zero and a standard deviation of one. The teacher baseline performance is based on changes in teacher-assigned student grades and is standardized with a mean of zero and a standard deviation of one. Significance levels: *** p<.01, ** p<.05, * p<.1.*

Table 3: The effect of unannounced public praise on standardized exam performance

	Exam	**Exam**
(β_1) Treatment	-0.055	-0.089
	(0.051)	(0.056)
	[0.322]	[0.184]
	{0.401}	{0.243}
(β_2) Top performer		0.054
		(0.052)
		[0.322]
		{0.950}
(β_3) Treatment * Top performer		0.258***
		(0.094)
		[0.038]
		{0.034}
Student Controls	yes	yes
Teacher controls	yes	yes
School Controls	yes	yes
N	6,639	6,639
F-value	349.27	214.38
R-squared	0.486	0.492

*Notes: The analysis is performed at the student level. The dependent variable is the student's exam performance, expressed in standard deviations. OLS regressions control for baseline student performance, student gender, subject fixed effects, profile type fixed effects, degree of urbanization, being a publicly-funded school, school size, baseline student and teacher quality at the school level, past year exam performance at the school level, and region fixed effects. In parentheses, heteroskedasticity robust standard errors are presented, clustered at the school level. In brackets, p-values are reported estimated using the wild bootstrap procedure suggested by Cameron et al. (2008), by clustering standard errors at the school level. Since the number of clusters is small, the more conservative Webb weights are used (Webb, 2013), implemented in Stata using the "boottest" estimator developed by Roodman et al. (2019), with 1000 replications. In braces, permutation-based p-values are reported, calculated by repeatedly redoing the random assignment, including the stratification, with 1000 replications, and by clustering the standard errors at the level of the school. The procedure is implemented in Stata using the "ritest" command (Heß, 2017). Significance levels: *** p<.01, ** p<.05, * p<.1.*

Table 4: Difference-in-difference analysis on the correlation between changes in teacher-assigned grades and standardized exam performance

	Correlation (g_{it}, $exam_{ij}$)
Post-intervention period	-0.068**
	(0.030)
	[0.026]
	{0.338}
Treatment	-0.006
	(0.051)
	[0.911]
	{0.912}
Post-intervention period * Treatment	0.014
	(0.055)
	[0.802]
	{0.803}
Student controls	yes
Teacher controls	yes
School controls	yes
N	497
F-value	17.60
R-squared	0.080

*Notes: The analysis is performed at the teacher level. The dependent variable is the correlation coefficient, at the teacher level, between the student's exam performance (expressed in standard deviations) and the student's pre-intervention changes in teacher-assigned grades (expressed in standard deviations). OLS regression controls for average baseline performance of a teacher's students, student gender composition, subject fixed effects, degree of urbanization, being a publicly-funded school, school size, baseline student and teacher quality at the school level, past year exam performance at the school level, and region fixed effects. In parentheses, heteroskedasticity robust standard errors are presented, clustered at the school level. In brackets, p-values are reported estimated using the wild bootstrap procedure suggested by Cameron et al. (2008), by clustering standard errors at the school level. Since the number of clusters is small, the more conservative Webb weights are used (Webb, 2013), implemented in Stata using the "boottest" estimator developed by Roodman et al. (2019), with 1000 replications. In braces, permutation-based p-values are reported, calculated by repeatedly re-doing the random assignment, including the stratification, with 1000 replications, and by clustering the standard errors at the level of the school. The procedure is implemented in Stata using the "ritest" command (Heß, 2017). Significance levels: *** p<.01, ** p<.05, * p<.1.*

Table 5: The effect of unannounced public praise on changes in teacher-assigned grades

	G_{it+1}	G_{it+1}
(α_1) Treatment	-0.150	-0.303**
	(0.115)	(0.121)
	[0.217]	[0.029]
	{0.252}	{0.038}
(α_3) Treatment * Top performer		0.528**
		(0.233)
		[0.038]
		{0.037}
Teacher Fixed Effects	yes	yes
Time Fixed Effects	yes	yes
N	821	821
F-value	11.75	169.61

*Notes: The analysis is performed at the teacher level. The dependent variable is the change in teacher-assigned grades calculated two months after the first intervention, expressed in standard deviations. In parentheses, heteroskedasticity robust standard errors are presented, clustered at the school level. In brackets, p-values are reported estimated using the wild bootstrap procedure suggested by Cameron et al. (2008), by clustering standard errors at the school level. Since the number of clusters is small, the more conservative Webb weights are used (Webb, 2013), implemented in Stata using the "boottest" estimator developed by Roodman et al. (2019), with 1000 replications. In braces, permutation-based p-values are reported, calculated by repeatedly re-doing the random assignment, including the stratification, with 1000 replications, and by clustering the standard errors at the level of the school. The procedure is implemented in Stata using the "ritest" command (Heß, 2017). Significance levels: *** p<.01, ** p<.05, * p<.1.*

Table A.1: Descriptive Statistics

Variable	Mean	Standard Deviation	Sample Size
School characteristics			
School Size	31.05	13.25	821
Share urban areas	0.91	0.29	821
Share publicly funded	0.85	0.36	821
Region			
North-East	0.52	0.50	821
South	0.34	0.48	821
North-West	0.14	0.34	821
Past year exam performance	8.19	0.92	821
Teacher characteristics			
Baseline Teacher Performance	0.15	0.69	821
% Teaching Subject			
Biology	0.08	0.26	821
Chemistry	0.05	0.23	821
English	0.17	0.37	821
Physics	0.10	0.30	821
Geography	0.08	0.27	821
Computer Science	0.07	0.26	821
History	0.08	0.26	821
Mathematics	0.17	0.38	821
Romanian Language	0.20	0.40	821
Student characteristics			
Female	0.50	0.50	88,612
Skipped classes	3.58	7.02	127,157
Baseline Student Performance	7.95	2.04	101,139
Final year profile			
Secondary School	0.27	0.44	6,639
Humanities (High-school)	0.19	0.40	6,639
Math (High-school)	0.27	0.44	6,639
Science (High-school)	0.17	0.37	6,639
Technical (High-school)	0.10	0.30	6,639
Exam grade	8.20	1.58	6,639

Table A.4.: Relationship between current learning and previous learning

	Pre-treatment *G*
Last year's *G*	0.079
	(0.060)
	[0.069*]
Student controls	yes
Teacher controls	yes
School controls	
N	371
F-value	4.39
R-squared	0.19

*Notes: The analysis is performed at the teacher level. The dependent variable is pre-intervention changes in teacher-assigned grades expressed in standard deviations. Last year's changes in teacher-assigned grades is also expressed in standard deviations. OLS regression controls for student gender composition, average baseline performance of a teacher's students, subject fixed effects, degree of urbanization, being a publicly-funded school, school size, baseline student and teacher quality at the school level, and region fixed effects. In parentheses, heteroskedasticity robust standard errors are presented. In brackets, p-values are reported estimated using the wild bootstrap procedure suggested by Cameron et al. (2008), by clustering standard errors at the school level. Since the number of clusters is small, the more conservative Webb weights are used (Webb, 2013), implemented in Stata using the "boottest" estimator developed by Roodman et al. (2019), with 1000 replications. Significance levels: *** p<.01, ** p<.05, * p<.1.*

Table A.5: Relationship between pre-intervention changes in teacher-assigned grades and exam performance

	Exam
Pre-intervention *G*	0.317***
	(0.030)
	[$<$ 0.001]
Student controls	yes
Teacher controls	yes
School controls	yes
N	5,308
F-value	483.92
R-squared	0.557

*Notes: The analysis is performed at the student level. The dependent variable is the student's exam performance, expressed in standard deviations. The pre-intervention changes in teacher-assigned grades are also expressed in standard deviations. OLS regression controls for baseline student performance, student gender, subject fixed effects, class profile fixed effects, degree of urbanization, being a publicly-funded school, school size, baseline student and teacher quality at the school level, past year exam performance at the school level, and region fixed effects. In parentheses, heteroskedasticity robust standard errors are presented, clustered at the school level. In brackets, p-values are reported estimated using the wild bootstrap procedure suggested by Cameron et al. (2008), by clustering standard errors at the school level. Since the number of clusters is small, the more conservative Webb weights are used (Webb, 2013), implemented in Stata using the "boottest" estimator developed by Roodman et al. (2019), with 1000 replications. Significance levels: *** p<.01, ** p<.05, * p<.1.*

Table A.6: Balance test for joint mean differences between the 'Attrition' and 'No attrition' group, each round

Round 1	No attrition	Attrition
Treatment	0.55	0.44
Top performer	0.25	0.32
Treatment * Top performer	0.14	0.15
N	821	34
F-value joint difference	1.15	
P-value joint difference	0.33	
Round 2	No attrition	Attrition
Treatment	0.55	0.54
Top performer	0.26	0.21
Treatment * Top performer	0.12	0.14
N	758	63
F-value joint difference	1.53	
P-value joint difference	0.21	
Round 3	No attrition	Attrition
Treatment	0.54	0.76
Top performer	0.24	0.17
Treatment * Top performer	0.12	0.14
N	729	29
F-value joint difference	1.93	
P-value joint difference	0.12	

*Notes: Balance tests are at the teacher level. Columns show mean differences between the 'No attrition' and the 'Attrition' groups. Significance levels: *** p<.01, ** p<.05, * p<.1.*

Table A.7: Balance tests at the teacher level

Variable	**Control mean**	**Treatment mean**	**P-value difference**
Student baseline perf.	-0.160	0.132	0.291
	(0.200)	(0.185)	
Teacher baseline perf.	0.006	-0.005	0.889
	(0.057)	(0.052)	
School size (no. teachers)	30.131	31.182	0.844
	(3.878)	(3.591)	
% Urban schools	0.903	0.916	0.911
	(0.083)	(0.077)	
% Publicly funded	0.908	0.794	0.425
	(0.104)	(0.096)	
% Female students	0.497	0.538	0.345
	(0.032)	(0.029)	
No. skipped classes	0.772	0.757	0.930
	(0.122)	(0.113)	
N	381	462	

*Notes: The balance tests are performed at the teacher level. The first two columns show variable means between the control group of schools, and the treated group of schools. In brackets, standard deviations are presented. The third column shows the p-values from two-sample t-tests on the null hypothesis that group means are equal. Student baseline performance and teacher baseline performance are standardized, with a mean of zero and a standard deviation of one. Standard errors are clustered at the school level. Significance levels: *** p<.01, ** p<.05, * p<.1.*

Table A.8: Differences between final year students and non-final year students

	Is a final year student			
Variable	**Coefficient**	**Standard Error**	**P-value**	**Bootstrapped P-value**
Baseline performance	0.011	0.008	0.172	0.184
Pre-treatment *G*	0.007	0.006	0.264	0.298
Female student	-0.011	0.006	0.352	0.386
In treated group	0.009	0.026	0.745	0.804
Urban school	0.030	0.031	0.336	0.460
Private funding	-0.089***	0.030	0.005	0.037
Randomization variables				
School size	-0.001	0.001	0.493	0.608
Baseline teacher performance	-0.091	0.075	0.234	0.302
Baseline student performance	0.013	0.016	0.392	0.451
F- value	2.49			
P-value	0.024			
R-squared	0.004			
N	48,101			

*Notes: The balance tests are performed at the student level. The dependent variable is an indicator taking value 1 if a student is in their final year (aged 14 and 18) and 0 otherwise. Student baseline performance and pre-treatment changes in teacher-assigned grades ("Pre-treatment G") are expressed in standard deviations. In the first column, coefficients from an OLS regression of the dependent variable on controls (β) are presented. In the second column, heteroskedasticity robust standard errors (σ) are presented, clustered at the school level. In the third column, the associated p-value on each coefficient is displayed. In the final column, p-values are reported estimated using the wild bootstrap procedure suggested by Cameron et al. (2008), by clustering standard errors at the school level. Since the number of clusters is small, the more conservative Webb weights are used (Webb, 2013), implemented in Stata using the "boottest" estimator developed by Roodman et al. (2019), with 1000 replications. Significance levels: *** p<.01, ** p<.05, * p<.1.*

Table A.9: The effect of unannounced public praise on standardized exam performance - teacher-level regressions

	Exam	**Exam**	**Exam**
Exam			
(β_1) Treatment	-0.088	-0.062	-0.095
	(0.069)	(0.073)	(0.089)
(β_2) Top performer	-0.005	0.039	-0.022
	(0.069)	(0.063)	(0.085)
(β_3) Treatment * Top performer	0.175*	0.265**	0.231
	(0.094)	(0.123)	(0.169)
Teacher controls	yes	yes	yes
School Controls	yes	yes	yes
N	305	154	175
F-value	110.07	369.56	59.53
R-squared	0.71	0.78	0.74

*Notes: The dependent variable is the average exam performance of a teacher's students, expressed in standard deviations. The first column uses the full teacher sample. The second column drops teachers who teach a small number of students (below the median). The third column only keeps teachers of compulsory exams subjects (Mathematics and Romanian Language) which by design have a larger number of students taking the final exam than teachers of elective subjects. OLS regressions control for average baseline student performance, average student gender, subject fixed effects, profile type fixed effects, degree of urbanization, being a publicly-funded school, school size, baseline student and teacher quality at the school level, past year exam performance at the school level, and region fixed effects. In parentheses, heteroskedasticity robust standard errors are presented, clustered at the school level. Significance levels: *** p<.01, ** p<.05, * p<.1.*

Table A.10: The effect of unannounced public praise on standardized exam performance with the full set of coefficients

	Exam	**Exam**
(β_1) Treatment	-0.055	-0.089
	(0.051)	(0.056)
	[0.322]	[0.184]
(β_2) Top performer		0.054
		(0.052)
		[0.322]
(β_3) Treatment * Top performer		0.258***
		(0.094)
		[0.038]
School controls		
Past year exam performance	0.542***	0.545***
	(0.108)	(0.116)
	[0.015]	[0.021]
In urban area	0.284**	0.295**
	(0.136)	(0.135)
	[0.150]	[0.119]
Private funding	-0.097	-0.127
	(0.134)	(0.142)
	[0.575]	[0.495]
South region	-0.126	-0.151
	(0.127)	(0.131)
	[0.473]	[0.409]
West region	-0.177	-0.198
	(0.120)	(0.129)
	[0.583]	[0.586]
Teacher controls		
(Reference: Teaches Biology)		
Teaches Chemistry	-0.294**	-0.281*
	(0.143)	(0.143)
	[0.068]	[0.088]
Teaches Physics	-0.459***	-0.479***
	(0.116)	(0.113)
	[0.011]	[0.013]
Teaches Geography	0.441***	0.451***
	(0.128)	(0.119)
	[0.003]	[0.001]
Teaches Computer Science	-0.415***	-0.451***

	(0.130)	(0.131)
	[0.006]	[0.007]
Teaches History	0.020	0.017
	(0.136)	(0.129)
	[0.895]	[0.903]
Teaches Math	-0.419***	-0.403***
	(0.123)	(0.125)
	[0.005]	[0.005]
Teaches Romanian language	-0.177*	-0.189*
	(0.102)	(0.096)
	[0.135]	[0.083]
Student controls		
Student baseline performance	0.161***	0.168***
	(0.015)	(0.014)
	[< 0.001]	[< 0.001]
Female	0.092***	0.086
	(0.025)	(0.025)
	[0.004]	[0.005]
Humanities profile	-0.322**	-0.310**
	(0.120)	(0.118)
	[0.003]	[0.004]
Math profile	0.055	0.055
	(0.082)	(0.081)
	[0.531]	[0.525]
Science profile	0.082	0.095
	(0.107)	(0.108)
	[0.490]	[0.417]
Technical profile	-0.501***	-0.511***
	(0.172)	(0.171)
	[0.017]	[0.015]
Randomization variables		
Baseline student quality in school	0.058	0.058
	(0.083)	(0.089)
	[0.582]	[0.621]
Baseline teacher quality in school	0.616***	0.545**
	(0.207)	(0.217)
	[0.033]	[0.039]
School size	-0.005	-0.006
	(0.004)	(0.004)
	[0.402]	[0.311]
N	6,639	6,639

F-value	349.27	214.38
R-squared	0.486	0.492

*Notes: The analysis is performed at the student level. The dependent variable is the student's exam performance, expressed in standard deviations. OLS regressions control for baseline student performance, student gender, subject fixed effects, profile type fixed effects, degree of urbanization, being a publicly-funded school, school size, baseline student and teacher quality at the school level, past year exam performance at the school level, and region fixed effects. In parentheses, heteroskedasticity robust standard errors are presented, clustered at the school level. In brackets, p-values are reported estimated using the wild bootstrap procedure suggested by Cameron et al. (2008), by clustering standard errors at the school level. Since the number of clusters is small, the more conservative Webb weights are used (Webb, 2013), implemented in Stata using the "boottest" estimator developed by Roodman et al. (2019), with 1000 replications. Significance levels: *** p<.01, ** p<.05, * p<.1.*

Table A.11: The effect of repeated public praise on standardized exam performance

	Exam
(ϕ_1) Treatment	0.006
	(0.079)
	[0.945]
($\phi_{3,1}$) Treatment * Top performer only once	-0.176
	(0.144)
	[0.345]
($\phi_{3,2}$) Treatment * Top performer more than once	0.338***
	(0.082)
	[0.007]
Student Controls	yes
School Controls	yes
N	6,639
F-value	587.47
R-squared	0.493

*Notes: The analysis is performed at the student level. The dependent variable is student exam performance, expressed in standard deviations. OLS regressions control for baseline student performance, student gender, subject fixed effects, class profile fixed effects, degree of urbanization, being a publicly-funded school, school size, baseline student and teacher quality at the school level, past year exam performance at the school level, and region fixed effects. In parentheses, heteroskedasticity robust standard errors are presented, clustered at the school level. In brackets, p-values are reported estimated using the wild bootstrap procedure suggested by Cameron et al. (2008), by clustering standard errors at the school level. Since the number of clusters is small, the more conservative Webb weights are used (Webb, 2013), implemented in Stata using the "boottest" estimator developed by Roodman et al. (2019), with 1000 replications. Significance levels: *** p<.01, ** p<.05, * p<.1.*

Table A.12.1: Changes in teacher-assigned grades throughout the intervention

	Treatment			Control		
	Mean	**Standard Deviation**	**N**	**Mean**	**Standard Deviation**	**N**
Pre-treatment	0.003	0.907	466	-0.004	1.103	389
Post unannounced praise	-0.073	0.977	452	0.095	1.013	369
Post announced praise 2	-0.099	0.789	418	-0.041	1.015	340
Post announced praise 3	-0.120	0.807	396	-0.060	0.935	333

Notes: Averages are at the teacher level. Columns show the average changes in teacher-assigned grades in the treatment and the control group, throughout the intervention. Changes in teacher-assigned student grades are standardized, with a mean of zero and a standard deviation of one.

Table A.12.2: The effects of unannounced and announced public praise on changes in teacher-assigned grades

	New G
($\gamma_{1,1}$) Treatment Round 1	-0.232**
	(0.113)
	[0.070]
($\gamma_{1,2}$) Treatment Round 2	-0.043
	(0.112)
	[0.722]
($\gamma_{1,3}$) Treatment Round 3	-0.069
	(0.123)
	[0.600]
($\gamma_{3,1}$) Treatment * Top performer Round 1	0.291**
	(0.143)
	[0.057]
($\gamma_{3,2}$) Treatment * Top performer Round 2	-0.327
	(0.195)
	[0.129]
($\gamma_{3,3}$) Treatment * Top performer Round 3	0.127
	(0.269)
	[0.669]
Teacher Fixed Effects	yes
Time Fixed Effects	yes
N	821
F-value	56.47

*Notes: The analysis is performed at the teacher level. The dependent variable is the changes in teacher-assigned grades calculated two months after the previous intervention, expressed in standard deviations. In parentheses, heteroskedasticity robust standard errors are presented, clustered at the school level. In brackets, p-values are reported estimated using the wild bootstrap procedure suggested by Cameron et al. (2008), by clustering standard errors at the school level. Since the number of clusters is small, the more conservative Webb weights are used (Webb, 2013), implemented in Stata using the "boottest" estimator developed by Roodman et al. (2019), with 1000 replications. Significance levels: *** p<.01, ** p<.05, * p<.1.*

Table A.12.3: The effects of repeated public praise on changes in teacher-assigned grades

	New *G*	**New** *G*
(δ_1) Treatment	-0.079	-0.115
	(0.086)	(0.089)
	[0.368]	[0.239]
($\delta_{3,1}$) Treatment * Top performer first time		0.037
		(0.109)
		[0.736]
($\delta_{3,2}$) Treatment * Top performer repeated time		-0.229
		(0.353)
		[0.575]
Teacher Fixed Effects	yes	yes
Time Fixed Effects	yes	yes
N	821	821
F-value	55.72	43.32

*Notes: The analysis is performed at the teacher level, over the entire academic year. The dependent variable is the changes in teacher-assigned grades calculated two months after the previous intervention, expressed in standard deviations. In parentheses, heteroskedasticity robust standard errors are presented, clustered at the school level. In brackets, p-values are reported estimated using the wild bootstrap procedure suggested by Cameron et al. (2008), by clustering standard errors at the school level. Since the number of clusters is small, the more conservative Webb weights are used (Webb, 2013), implemented in Stata using the "boottest" estimator developed by Roodman et al. (2019), with 1000 replications. Significance levels: *** p<.01, ** p<.05, * p<.1.*

Table A.12.4 : Treatment effects throughout all periods

	New *G* (2 rounds)	**New** *G* (3 rounds)
(ψ_1) **Treatment**	-0.249**	-0.265**
	(0.099)	(0.104)
	[0.030]	[0.026]
Treatment * Type		
(ψ_2)T*Top1	0.473**	0.490**
	(0.215)	(0.221)
	[0.049]	[0.044]
(ψ_3)T*(NTop2 & Top1)	0.298	0.314
	(0.279)	(0.278)
	[0.304]	[0.272]
(ψ_4)T*(Top2 & NTop1)	-0.124	-0.116
	(0.524)	(0.254)
	[0.642]	[0.656]
(ψ_5)T*(Top2 & Top1)	-0.068	-0.051
	(0.445)	(0.449)
	[0.886]	[0.915]
(ψ_6)T*(NTop3 & NTop2 & Top1)		0.525*
		(0.301)
		[0.122]
(ψ_7)T*(NTop3 & Top2 & NTop1)		0.179
		(0.261)
		[0.523]
(ψ_8)T*(NTop3 & Top2 & Top1)		0.401
		(0.483)
		[0.472]
(ψ_9)T*(Top3 & NTop2 & NTop1)		0.106
		(0.196)
		[0.576]
(ψ_{10})T*(Top3 & NTop2 & Top1)		0.606
		(0.536)
		[0.318]
(ψ_{11})T*(Top3 & Top2 & NTop1)		-0.625
		(0.490)
		[0.259]
(ψ_{12})T*(Top3 & Top2 & Top1)		-0.455
		(1.028)
		[0.696]

Type Fixed Effects	yes	yes
Teacher Fixed Effects	yes	yes
Time Fixed Effects	yes	yes
N	821	821
F-value	88.53	122.81

*Notes: The analysis is performed at the teacher level. The dependent variable is the changes in teacher-assigned grades calculated two months after the previous intervention, expressed in standard deviations. In parentheses, heteroskedasticity robust standard errors are presented, clustered at the school level. In brackets, p-values are reported estimated using the wild bootstrap procedure suggested by Cameron et al. (2008), by clustering standard errors at the school level. Since the number of clusters is small, the more conservative Webb weights are used (Webb, 2013), implemented in Stata using the "boottest" estimator developed by Roodman et al. (2019), with 1000 replications. Significance levels: *** p<.01, ** p<.05, * p<.1.*

Table A.13: The effect of unannounced public praise on changes in teacher-assigned grades, for different quantiles of the performance distribution

	New *G*
Treatment	-0.374**
	(0.160)
	[0.027]
Treatment * Top performers (Quantile 1)	0.599**
	(0.251)
	[0.033]
Treatment * Quantile 2	0.199
	(0.161)
	[0.224]
Treatment * Quantile 3	0.154
	(0.220)
	[0.508]
Teacher Fixed Effects	yes
Time Fixed Effects	yes
N	821
F-value	152.71

*Notes: The analysis is performed at the teacher level. The dependent variable is the changes in teacher-assigned grades calculated two months after the first intervention, expressed in standard deviations. In parentheses, heteroskedasticity robust standard errors are presented, clustered at the school level. In brackets, p-values are reported estimated using the wild bootstrap procedure suggested by Cameron et al. (2008), by clustering standard errors at the school level. Since the number of clusters is small, the more conservative Webb weights are used (Webb, 2013), implemented in Stata using the "boottest" estimator developed by Roodman et al. (2019), with 1000 replications. Significance levels: *** p<.01, ** p<.05, * p<.1.*

Table A.14.1: The effect of public praise on student attendance

	Increase Skipped Classes (2 Months After Intervention)	**Increase Skipped Classes** (End of Academic Year)
(α_1) Treatment	0.030	-0.078
	(0.083)	(0.074)
	[0.775]	[0.373]
	{0.775}	{0.365}
(α_2) Top performer	0.058	-0.084
	(0.056)	(0.053)
	[0.390]	[0.249]
	{0.261}	{0.679}
(α_3) Treatment * Top performer	-0.087	-0.066
	(0.077)	(0.081)
	[0.360]	[0.494]
	{0.370}	{0.531}
Student Controls	yes	yes
Teacher controls	yes	yes
School Controls	yes	yes
N	101,021	101,021
F-value	4.45	25.57
R-squared	0.01	0.07

*Notes: The analysis is performed at the student level. The dependent variable is the increase in student skipped classes, compared to the number of skipped classes prior to the intervention, standardized with a mean of zero and a standard deviation of one. The first column shows the increase in skipped classes two months after the intervention. The second column shows the increase in skipped classes by the end of the academic year. OLS regressions control for baseline student performance, student gender, subject fixed effects, degree of urbanization, being a publicly-funded school, school size, baseline student and teacher quality at the school level, and region fixed effects. In parentheses, heteroskedasticity robust standard errors are presented, clustered at the school level. In brackets, p-values are reported estimated using the wild bootstrap procedure suggested by Cameron et al. (2008), by clustering standard errors at the school level. Since the number of clusters is small, the more conservative Webb weights are used (Webb, 2013), implemented using the boottest estimator developed by Roodman et al. (2019), with 1000 replications. In braces, permutation-based p-values are reported, calculated by repeatedly re-doing the random assignment, including the stratification, with 1000 replications, and by clustering the standard errors at the level of the school. The procedure is implemented in Stata using the "ritest" command (Heß, 2017). Significance levels: *** p<.01, ** p<.05, * p<.1.*

Table A.14.2: The effect of public praise on student attendance, by student baseline performance

	Low performing students		**High performing students**	
	Increase Absences	**Increase Absences**	**Increase Absences**	**Increase Absences**
	(2 Months After)	(End Year)	(2 Months After)	(End Year)
(α_1) Treatment	0.007	-0.065	0.048	-0.100
	(0.083)	(0.127)	(0.089)	(0.049)
	[0.941]	[0.688]	[0.711]	[0.077*]
	{0.962}	{0.650}	{0.723}	{0.107}
(α_2) Top performer	0.055	-0.113	0.041	-0.064
	(0.064)	(0.091)	(0.047)	(0.033)
	[0.471]	[0.345]	[0.416]	[0.156]
	{0.392}	{0.794}	{0.401}	{0.380}
(α_3) Treatment * Top performer	-0.101	-0.178	-0.061	0.012
	(0.074)	(0.124)	(0.074)	(0.053)
	[0.237]	[0.228]	[0.482]	[0.810]
	{0.427}	{0.283}	{0.490}	{0.860}
Student Controls	yes	yes	yes	yes
Teacher controls	yes	yes	yes	yes
School Controls	yes	yes	yes	yes
N	35,323	35,323	65,698	65,698
F-value	11.01	25.67	11.89	48.61
R-squared	0.01	0.07	0.01	0.06

*Notes: The analysis is performed at the student level. The dependent variable is the increase in student skipped classes, compared to the number of skipped classes prior to the intervention, standardized with a mean of zero and a standard deviation of one. The first and third columns show the increase in skipped classes two months after the intervention. The second and fourth columns show the increase in skipped classes by the end of the academic year. OLS regressions control for baseline student performance, student gender, subject fixed effects, degree of urbanization, being a publicly-funded school, school size, baseline student and teacher quality at the school level, and region fixed effects. In parentheses, heteroskedasticity robust standard errors are presented, clustered at the school level. In brackets, p-values are reported estimated using the wild bootstrap procedure suggested by Cameron et al. (2008), by clustering standard errors at the school level. Since the number of clusters is small, the more conservative Webb weights are used (Webb, 2013), implemented using the boottest estimator developed by Roodman et al. (2019), with 1000 replications. In braces, permutation-based p-values are reported, calculated by repeatedly re-doing the random assignment, including the stratification, with 1000 replications, and by clustering the standard errors at the level of the school. The procedure is implemented in Stata using the "ritest" command (Heß, 2017). Significance levels: *** p<.01, ** p<.05, * p<.1.*

Table A.15: The effect of unannounced public praise on standardized exam performance with propensity scores as controls

	Exam	**Exam**
(β_1) **Treatment**	-0.057	-0.085
	(0.053)	(0.059)
	[0.392]	[0.261]
	{0.406}	{0.316}
(β_2) **Top performer**		0.058
		(0.053)
		[0.294]
		{0.287}
(β_3) **Treatment * Top performer**		0.222**
		(0.097)
		[0.087]
		{0.073}
Student Controls	yes	yes
Teacher controls	yes	yes
School Controls	yes	yes
Treatment Assignment Propensity Score	yes	yes
N	6,362	6,362
F-value	619.11	495.14
R-squared	0.49	0.50

*Notes: The analysis is performed at the student level. The dependent variable is the student's exam performance, expressed in standard deviations. OLS regressions control for baseline student performance, student gender, subject fixed effects, profile type fixed effects, degree of urbanization, being a publicly-funded school, school size, baseline student and teacher quality at the school level, past year exam performance at the school level, and region fixed effects. In parentheses, heteroskedasticity robust standard errors are presented, clustered at the school level. In brackets, p-values are reported estimated using the wild bootstrap procedure suggested by Cameron et al. (2008), by clustering standard errors at the school level. Since the number of clusters is small, the more conservative Webb weights are used (Webb, 2013), implemented in Stata using the "boottest" estimator developed by Roodman et al. (2019), with 1000 replications. In braces, permutation-based p-values are reported, calculated by repeatedly re-doing the random assignment, including the stratification, with 1000 replications, and by clustering the standard errors at the level of the school. The procedure is implemented in Stata using the "ritest" command (Heß, 2017). Significance levels: *** p<.01, ** p<.05, * p<.1.*

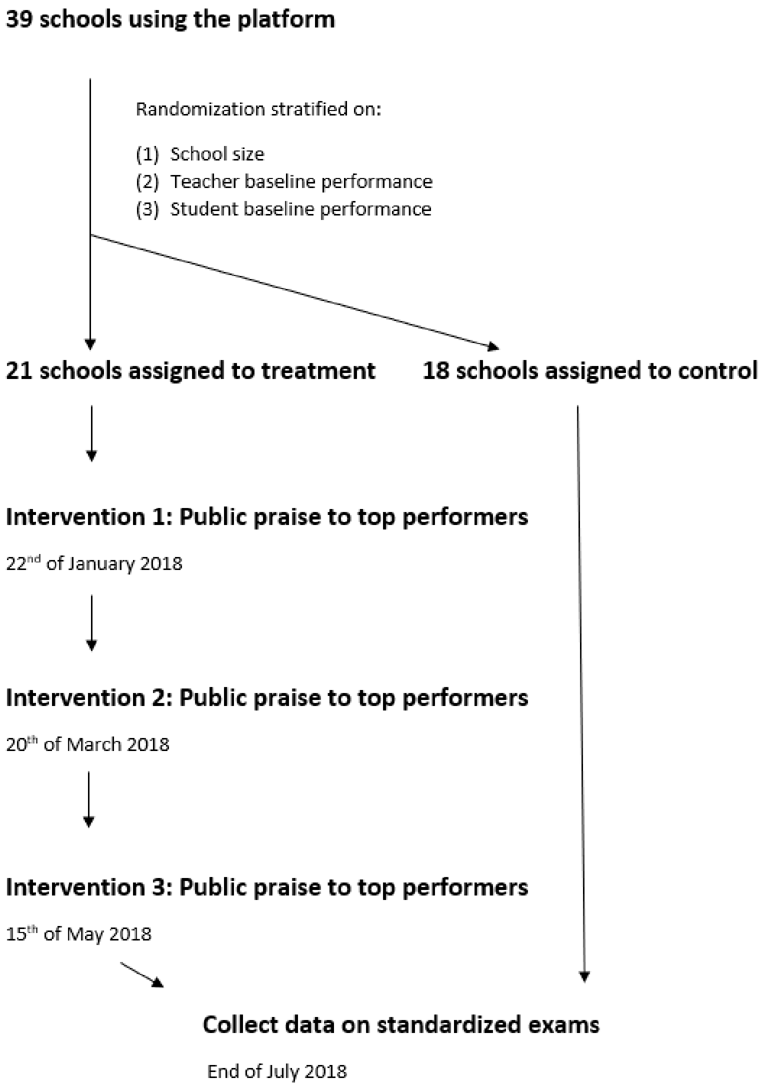

39 schools using the platform
Randomization stratified on:
(1) School size
(2) Teacher baseline performance
(3) Student baseline performance
21 schools assigned to treatment
18 schools assigned to control
Intervention 1: Public praise to top performers
22nd of January 2018
Intervention 2: Public praise to top performers
20th of March 2018
Intervention 3: Public praise to top performers
15th of May 2018
Collect data on standardized exams
End of July 2018

Figure A.2. : Schematic representation of the intervention time-line

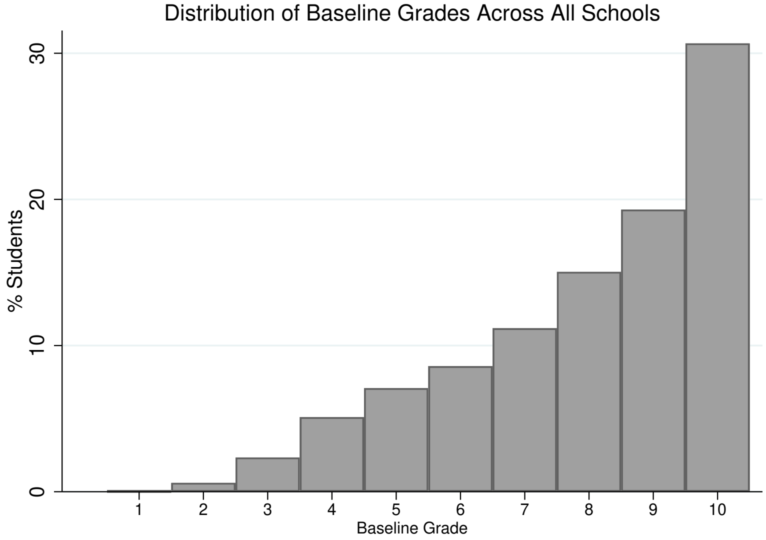


Figure A.3. : Distribution of Baseline Grades

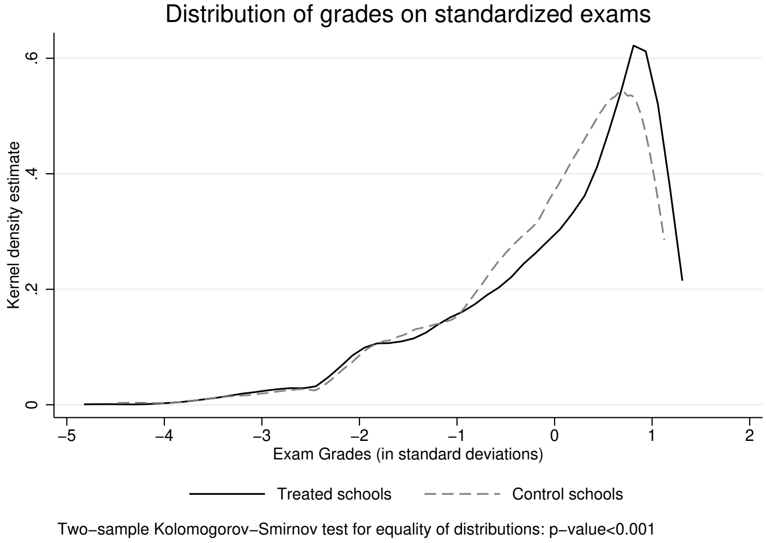
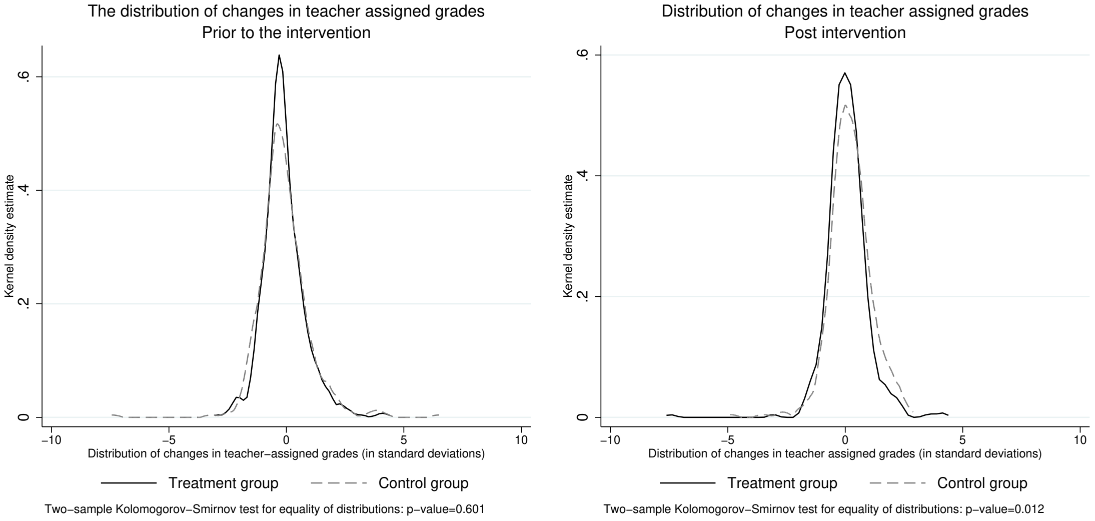


Figure A.4. : Distribution of exam grades by treatment status


The distribution of changes in teacher assigned grades
Prior to the intervention

Distribution of changes in teacher assigned grades
Post intervention

Two−sample Kolomogorov−Smirnov test for equality of distributions: p−value=0.601

Two−sample Kolomogorov−Smirnov test for equality of distributions: p−value=0.012

Figure A.5. : Distribution of teacher-assigned grades by treatment status