Effects of Timing and Reference Frame of Feedback: Evidence from a Field Experiment in Secondary Schools[*]

Mira Fischer[†]and Valentin Wagner[‡]

19th May 2017

[PRELIMINARY DRAFT - PLEASE DO NOT QUOTE]

Abstract

We analyze the effectiveness of relative performance feedback in a high-stakes exam when the timing and the reference frame of feedback are manipulated. In a field experiment in secondary schools, students of grades 5 and 6 were provided with information about their absolute rank in the last exam, their change in ranks, or with no feedback. With respect to timing, students got their feedback 1-3 days or immediately before the last exam of the semester. Overall, any type of feedback—absolute and change—increased students' performance when it was given 1-3 days before the exam compared to immediate feedback and is most effective for students who recently suffered a decrease in their performance.

Keywords: Timing and reference frames of feedback, relative performance feedback, high-stakes testing, field experiment, rank

JEL Codes: D03, D83, J24, I21, C93

[*] We would like to thank Dirk Sliwka, Gerhard Riener, Armin Falk, Paul Heidhues, Hannah Schildberg-Hörisch, Bertil Tungodden, Alexander Cappelen, the participants of the 2nd Winter School on Applied Microeconomics, the 3rd Workshop on Educations Economics, as well as seminar participants at the University of Cologne, the University of Düsseldorf and the Choice Lab in Bergen for helpful comments and suggestions. Particular thanks to all the teachers, parents and pupils who participated in the experiment. This research has been conducted with approval of the ethics commission of the Heinrich-Heine-Universität Düsseldorf. The usual disclaimer applies.

[†] University of Cologne, Seminar of Personnel Economics and Human Resource Management, Albertus-Magnus-Platz, 50923 Köln, Germany, phone: +49 221 470-6311, email: mira.fischer@uni-koeln.de

[‡] University of Düsseldorf, Düsseldorf Institute for Competition Economics, Universitätsstr. 1, 40225 Düsseldorf, Germany, phone: +49 211 81-10239, email: wagner@dice.uni-duesseldorf.de

1 Introduction

Feedback about performance relative to one's colleagues is widely used in the workplace in addition to incentive pay or by itself, for example because the latter may be difficult to implement or socially unacceptable. In education there are strong concerns that material incentives may crowd out intrinsic motivation so grades assigned on a curve are used both as status incentives and as means of giving feedback about relative performance. Indeed, feedback about one's past performance is thought to be a powerful means to help humans improve their future performance [Thaler et al., 2013] and many experimental studies in both economics and psychology find that it "works" [Azmat and Iriberri, 2016, 2010, Tran and Zeckhauser, 2012]. However, it is also frequently found to backfire [Azmat et al., 2016, Bradler et al., 2016, Ashraf et al., 2014, Barankay, 2012]or to be ineffective [Eriksson et al., 2009].[1] Determining which design features make feedback successful or not is therefore an important object of study. The question of what makes feedback effective has received rather little attention from researchers although it is highly relevant in all contexts where the ability to motivate people is crucial, such as labor, education, or sports. Past research has found that feedback comparing different people's performance with each other, for example by revealing relative ranks, is more effective than feedback referring to an absolute standard [Azmat and Iriberri, 2010] and there are mixed findings about whether public and private rank feedback is more effective [Tran and Zeckhauser, 2012, Ashraf et al., 2014, Gill et al., 2016, Hannan et al., 2013, Tafkov, 2013]. Other important features of feedback, for example, how it is timed and how it makes use of social comparison, are potentially important determinants of whether it helps to improve performance but evidence on them is scarce.

Feedback is often not deliberately timed in the workplace or educational settings although in both settings different types of effort exerted at different times, e.g. preparation effort and effort at the task itself, may influence outcomes.[2] While earlier feedback might have a stronger impact on preparation efforts, feedback given more immediately before a task may potentially have a stronger effect on effort at the task itself. However, depending on the task at hand, preparation effort or effort at the task may be more important and thus for different tasks feedback should be timed differently. Moreover, feedback can be timed with respect to prior observations of performance. Although there might be a tendency to try to reinforce effort by giving feedback after observing high performance or personal improvement, it is not clear whether it actually is more motivating than feedback after low performance or slacking off. On the contrary, a warning shot might have a stronger positive effect on motivation than the award of laurels as the first might wake people up whereas they might be tempted to rest on the latter.

Adam Smith thought that "rank among our equals, is, perhaps, the strongest of all our desires" [Smith, 1759]. Empirical evidence suggests that people are, indeed, strongly motivated by ranks, even in the absence of any tangible benefits [Charness and Rabin, 2002], and are particularly motivated to achieve a first and avoid a last place in a ranking [Kuziemko et al., 2014, Gill et al., 2016]. However, rank feedback that compares

[1] See also [Kluger and DeNisi, 1998, 1996] for evidence in the psychological literature.

[2] See Levitt et al. [2016a], Wagner [2016] for studies disentangling the effort from the learning effect in educational settings.

one's level of performance to one's peers' levels does not properly capture individual improvement over time as the reference group might also be moving upward.[3] This feature of rank feedback might have downsides in education in particular where large differences in ability levels can often be found within the same class. If heterogeneity is large, revealing ability differences may reduce the motivation to exert effort in tournament settings. [Gürtler and Harbring, 2010]. Feedback that compares students not in terms of their levels but in terms of their changes in performance might help to mitigate this problem while maintaining the motivational effects of social comparison. Change feedback may also help to promote the belief that skills can be developed by exerting effort (also called a "growth mindset" in the psychological literature, see Paunesku et al. 2015, O'Rourke et al. 2014). For these reasons, feedback that captures individual changes in performance over time may better help students improve their academic skills than relative feedback that relies on making cross-sectional skill differences salient. As such it may motivate both weaker and stronger students to invest more in their skills. Academic skills are a strong determinant of a person's health [Cutler and Lleras-Muney, 2006], wealth [Hanushek et al., 2015, Oreopoulos, 2007] and well-being [Oreopoulos and Salvanes, 2011] and also have strong effects on society [Milligan et al., 2004] but even in developed countries a large proportion of people fail to acquire a minimum level of academic skills needed to participate in civil society, or to find employment . Because of a strong complementarities of skill formation at different stages of the education production function, gaps in academic skills at a younger age become wider as people age, which is why interventions should target younger students, if possible [Cunha and Heckman, 2007]. In light of these facts the question of whether feedback in schools can be improved by deliberately selecting its *references frame* (levels or changes in performance) as well as its *timing* is an important one.

In order to shed light on the effects of timing and reference frame of relative feedback on high-stakes educational outcomes, we study a field experiment in which students in 19 secondary school classes received private written feedback from their teachers. Both the reference frame and timing were randomized to allow for causal identification of effects. With regard to the timing, students received the feedback intervention either 1-3 days or immediately before they wrote the final mathematics exam of the school year, which immediately affected progression to the next grade. With regard to reference frame, students within the same class either received feedback about (i) their performance rank in the last exam, or (ii) feedback about their change in rank between the second last and the last exam, or (iii) no feedback.

We find that the effect of feedback on subsequent performance depends on the timing of feedback both with respect to the subsequent exam and with respect to prior performance but not on whether feedback is given in terms of changes or levels of performance. Both change and level feedback, when given 1-3 days prior to an exam, have an overall positive effect on performance. This effect is driven by giving feedback to those students who recently suffered a decrease in their performance. Feedback given to students immediately before the exam has no significant effect on performance.

[3] Whether students have or have not a preference of being socially ranked is unclear. However, Gill et al. [2016] show in a lab experiment that workers have a pure taste for being ranked which is independent of long-term reputational considerations or any desire for compensation.

Our experimental design allows us to make several contributions to the literature. To our knowledge this is the first study that varies the timing of feedback. It is also the first to compare two general types of relative feedback (in terms of levels and changes of performance). Furthermore, it tests written feedback on a sample of secondary school students (aged around 10-11 years), while so far researchers have exploited data of natural experiments or tested feedback on university students. Our results are directly relevant for educators but the general findings potentially extend to other settings where feedback is given with the intention to increase motivation, such as the workplace or sports.

The remainder of this paper is organized as follows. The next section gives a brief overview about the relevant literature. In section 3 we report on the results of our pretest. Section 4 describes our experimental procedures. Section 5 presents the results which are discussed in the section thereafter. Section 7 concludes.

2 Related Literature

Field experiments in education: incentives versus feedback Although the effectiveness of teachers is very heterogeneous, surprisingly little of it can be explained by observed teacher characteristics [Hanushek and Rivkin, 2006] which makes it hard to improve educational outcomes by screening for good teachers. Besides screening, economists traditionally focus on the introduction of incentives in order to raise productivity. In recent years, field experiments on monetary [Levitt et al., 2016b, Fryer et al., 2012, Bettinger, 2012, Fryer, 2013] and non-monetary [Levitt et al., 2016a, Jalava et al., 2015, Wagner and Riener, 2015] incentives for teachers and/or students have produced mixed results.[4] At the same time, other field experiments have tested the effects of feedback on educational outcomes. As compared to incentive interventions, feedback interventions have several advantages that make them attractive. First they have a reduced risk of crowding out intrinsic motivation or the effects of grade incentives that may already be present, second, they face fewer concerns by teachers and parents as feedback is widely used and accepted in an educational context, and third, feedback interventions can be virtually cost-free.

Most studies on feedback so far rely on university student samples and generally find that relative performance feedback boosts performance. Tran and Zeckhauser [2012] provide Vietnamese students participating in an English-testing experiment either with a private feedback (by phone) or private plus public feedback (postings on the university's noticeboard and website) about their ranking in the in-course mock exams. Overall, the authors find a positive effect of feedback on the final TOEIC test and that public plus private feedback tends to outperform purely private feedback. However, the difference was only marginally significant. A more recent study by Bandiera et al. [2015] exploit data of a natural experiment in the UK where some university students were provided with a private and absolute feedback on their past exam performance and others were not. Feedback on exam performance improved students' future performance mostly for more able students and for students who initially start with less information about the academic environment.

[4] Damgaard and Nielsen [2017] recently review the use of nudges and other behaviorally motivated interventions in education.

Azmat et al. [2016] provided feedback on their position in the grade distribution to college students every six months over a period of three years. They find that students who received feedback suffered a decrease in their performance relative to a control group. This effect is driven by students who underestimated their relative performance in the absence of feedback.

While the studies described above analyze the effect of feedback on performance among university students, we are aware of only one study on school aged children which exploits data from a *natural* field experiment [Azmat and Iriberri, 2010] and there is—to our knowledge—no *randomized controlled* field experiment on the effectiveness of relative performance feedback on high stakes testing outcomes on school children. Our paper is therefore most closely related to the paper by Azmat and Iriberri [2010] with respect to the population studied. The authors report on the motivational effect of relative performance feedback among high school students in Spain (grades 7 - 10). For one school year, a high school in the Basque Country adopted a new application to produce report cards giving students the information whether they were performing above (below) the class average as well as the distance from this average. Before and after this change, report cards informed students only about their own grade point average. The new relative performance feedback had positive effects and increased students' grades by 5 %. However, the effect disappeared as soon as the information was removed.

Different types of feedback The seminal papers by Mas and Moretti [2009] and Falk and Ichino [2006] show that indirect feedback, that is, observed performance of peers, increases a worker's performance even if better performance does not increases monetary payoffs.[5] This shows that pure peer effects or status-seeking effects might play a role when giving public feedback. A large number of studies give direct feedback about performance, in particular rank feedback, and rely on worker populations or laboratory experiments. They typically also find a positive increase in performance when participants learn their rank in the population. In line with Mas and Moretti [2009] and Falk and Ichino [2006], Charness et al. [2014], finds status-seeking concerns as participants in a laboratory experiment are willing to pay for sabotaging the performance of others—pay to reduce the score of competitors—when getting relative performance feedback. Kuhnen and Tymula [2012] show that performance increases and participants expect to rank better when told that they may privately learn their ranking, while Gill et al. [2016] find a u-shaped rank response function—first place loving and last place loathing—when participants in the laboratory privately or publicly learn their rank in a real effort task. Subjects increase their effort the most after being ranked first or last but these motivational effects do not depend on whether the feedback was reported publicly or privately. Azmat and Iriberri [2016] find in a laboratory experiment that relative performance feedback is only effective to increase performance when the payment depends on performance but not if subjects are payed a flat-rate.[6] Most studies report on how the provision of feedback per se affects performance but do not explicitly differentiate

[5] See also Dechenaux et al. [2015] for a summary of the findings in the tournament literature.

[6] The findings by Azmat and Iriberri [2016] seem to be contrary to Mas and Moretti [2009] and Falk and Ichino [2006]. However, it is important to notice that Azmat and Iriberri [2016] give solely private feedback and therefore differ from settings in which pure peer effects or status-seeking effects play a role as by Mas and Moretti [2009] and Falk and Ichino [2006].

between the type of feedback, i.e. whether it is positive or negative. However, Bradler et al. [2016] show that the type of feedback might matter as they find that only bad feedback—not receiving the award—motivates employees. Nevertheless, there is also evidence that rank feedback does not improve performance [Eriksson et al., 2009] or even backfires [Barankay, 2012, Hannan et al., 2008]. Barankay [2012] finds that sales people increase performance when rank feedback is removed and Hannan et al. [2008] show that relative performance feedback decreases performance of participants compensated under a tournament incentive scheme when the feedback is sufficiently precise.

3 Pretest

Students in grades five and six usually do not get private feedback about their relative rank within the classroom or about the relative change of their rank. If any, teachers provide their class with a statistic about the frequency of grades *after* the exam. However, this information only allows students to infer in which range their rank is but they do not learn their exact rank.[7] As rank feedback is uncommon in the German school system, we had to ensure that students understand the feedback information in order to know whether potentially null results are due to a lack of understanding or the ineffectiveness of providing feedback. To test the applicability of our feedback notes, we therefore conducted a survey before implementing the field experiment in 4 classes of 6 schools with a total of 151 pupils of the same age group. This was a convenience sample gathered through personal contacts.

The pretest consisted of a two-paged questionnaire.[8] On the first page students saw the feedback note of the fictional student "Paul" and were asked to imagine themselves in Pauls' position. On the back side, students had to shortly summarize in their own words the information of the first page, answer a questions about Pauls' emotional state after having read the feedback note, a question on Pauls' motivational level and questions that checked students' comprehension of the feedback. We also asked students whether they know their class size which is a crucial assumption to understand relative rank feedback. We randomly varied Pauls' feedback note and presented students either with a change feedback or a level feedback. Furthermore, we varied the size of the change in ranks (-6, -3, 0, 3, or 6) respectively the level in ranks (5, 15, or 25).

Overall students seem to understand the feedback notes. 85.56% of the students could correctly calculate how Pauls' rank changed and 94.74% could correctly determine the position of Pauls' rank in the Level-Feedback condition. Moreover, 86.09% of the students know the exact size of their class. The mean responses to the questions concerning Pauls' emotions and motivation are presented in Figure 1 distinguished by the reference frame of feedback. On average, students in the Change-Feedback seem to be more motivated than students in the Level-Feedback (3.87 vs 3.53, p = 0.0745) while the two references frames of feedback do not differently affect emotions (3.05 vs 2.80, p = 0.2353). However, the significant difference between the Change-

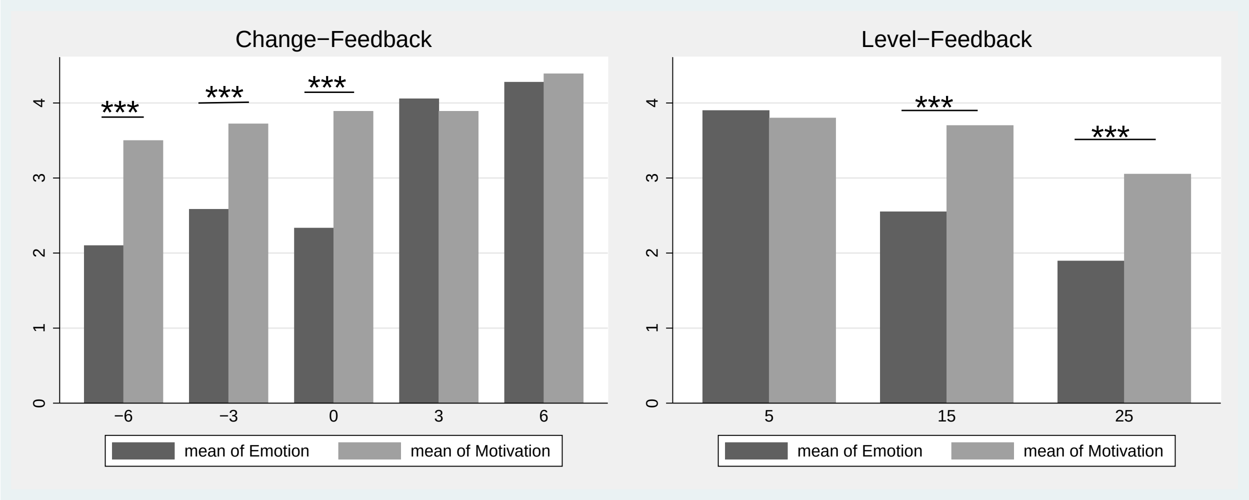
[7] If, for example, students learn that 5 students got the grade A, 5 the grade B, 5 the grade C and 5 the grade D, those students with a C know that their rank is between 11-15, but they do not now their rank within this group.

[8] See Appendix C.

and Level-Feedback in responses to motivation can only be interpreted as a tendency and not as definite as differences could be a result of the chosen ranks in the level feedback and might become insignificant when testing different ranks, e.g. 1, 10, 20 and 30. Analyzing mean emotions and mean motivation within the Change- and Level-Feedback, we observe a similar pattern. Emotions and motivation are rated higher with positive feedback. Interestingly, the difference between emotions and motivation—motivation scores higher than emotion—is large and significant for negative feedback but not for positive feedback. Furthermore, the mean level of motivation does not drop below 3 with negative feedback. On a 1 to 5 scale, this can be interpreted as a non-negative motivational effect of negative feedback.[9]

Finding from Pretest: *Negative feedback tends to have a unfavorable effect on the emotional state of students but does not seem to discourage students in terms of motivation.*

Figure 1: Pretest - Stated Emotions and Motivation by Reference Frame of Feedback



Note: This graph presents the answers of the pretest separately for the Change- (left) and Level-Feedback (right). Dark bars are mean responses to the question *How do you think does Paul feel after reading the note?*, gray bars are mean responses to the question *How much do you think is Paul motivated to exert effort in the upcoming math exam?*. Both are measured on a 1 (not at all) to 5 (very much) scale. Feedback notes in the pretest were varied such that students faced either a change in Pauls' rank of -6, -3, 0, 3 or 6 in the Change-Feedback or the ranks 5, 15 or 25 in the Level-Feedback. Differences between emotions and motivation are based on a mean-comparison tests.

4 Experimental Intervention

The experiment was conducted in 7 secondary schools with in total 19 classes in the cities of Bonn, Cologne and Düsseldorf, Germany and was approved by the ethics commission of the Heinrich-Heine-Universität

[9] We interpreted a mean score of around 2 or lower as a negative effect on motivation.

Düsseldorf. 352 students of grades five and six participated during May and June 2016.[10] Researchers were never present in the classroom to maintain a "natural" examination situation. Hence, the experiment was conducted solely by the teacher. To train teachers how to conduct the experiment, we visited the schools in the run-up of the experiment. During this meeting, the exact schedule and expiration of the experiment was described and teachers' questions were answered. In total teachers received two envelopes from us with necessary material to run the experiment. The first envelope contained written teacher instructions, consent forms to be signed by parents and a "list of grades" to be filled out by teachers. This list contained information on the number of points and the grade pupils got in the first and second exam of the semester. We prepared the personalized feedback notes by calculated students' rank and their change in ranks.[11] The feedback notes were folded to increase teachers' cost to look at these notes and students' names were written on the front.[12] As describes earlier, the comprehensibility of feedback notes were tested prior to the experiment in a pretest among students of the same age group. The second envelope was send in a timely manner close to the last exam of the semester (to further ensure that teachers have not the time to look at the feedback notes) containing the feedback notes, student questionnaires and instructions how to exactly distribute the feedback notes.

The student questionnaires consisted of 5 pages and was answered by students—depending on the treatment—either after reading the feedback notes or after the math exam.[13] The questionnaire asked for pupils' background characteristics and self-related beliefs. Questions on self-related beliefs are based on validated questionnaires and measured locus of control [taken from PISA and adjusted for age; based on Rotter, 1966], academic and math self-efficacy (taken from PISA, based on) and self-esteem (German version of the Rosenberg self-esteem scale,).

After students wrote the exam, teachers were required to send the grades of the exam as well as the student questionnaires to the researchers. Thereafter, we asked teachers to fill out an online teacher questionnaire.

Treatments We are interested in how feedback affects a students' performance in a high-stakes math exam. Based on a 2 X 3 design, we vary both the *timing* of feedback and the *reference frame* of feedback independently. The timing of feedback was varied on class-level whereas the content of feedback was varied on student-level. First, with respect to the timing of feedback, students received feedback about their past performance (their rank) 1-3 days before the last math exam in the school year in the "*Early-Feedback Treatment*" (Early). In the "*Late-Feedback Treatment*" (Late) students received the feedback immediate before the last exam—in the same testing hour, seconds before the teacher handed out the test questions.

[10] We contacted 142 secondary schools in North Rhine- Westphalia (NRW) by using a list of schools that is publicly available from the Ministry of Education of NRW. 23% of the schools responded and 39% (13 out 33) of these schools were generally interested in participating. After further consultation with schools, 7 schools finally participated.

[11] See Appendix D for a sample of the feedback notes.

[12] We did not put the feedback notes in closed envelopes to ensure that teachers do not look at them because this could have caused too much disturbance within the classroom and pupils might not open the envelope. Furthermore, we do not belief that teachers look at the feedback notes at this is time consuming and teachers usually try to avoid extra workload.

[13] Students who received the feedback notes 1-3 days before the exam answered the questionnaire after having read the feedback notes. Students who received the feedback notes immediately before the exam answered the questionnaire after the exam. In the latter, the questionnaire was a shorter version of the questionnaire answered by students receiving the feedback notes earlier.

Second, with respect to the reference frame of feedback, each student in all treatment groups received a folded and private feedback note which had to be returned to the teacher afterwards. The feedback note in the *Control Group* was a "good luck" note ("I wish you great success in your exam") but students did not get any information about their past performance. In the *Change Feedback* condition, students got information how their rank has changed between exam 1 and exam 2 ("Relative to your classmates, you improved/worsened your performance in the last math exam by XX places"). In this manner, students did not learn their absolute rank in both exams and more importantly students at the bottom of the rank distribution could receive a positive feedback. Students in the *Level Feedback* condition were informed about their relative rank in exam 2 ("Relative to your classmates, you achieved with your performance in the last math exam, the XX th place"). No feedback was provided on their rank in exam 1 and hence students did not get any information about the change in rank between exam 1 and exam 2.

5 Results

This section is organized as follows: first, we describe our randomization strategy and discuss concerns about non-random self-selection into treatment groups. Thereafter, we present our data and descriptive statistics before analyzing the impact of feedback on students' performance. We first examine the role of feedback timing and then examine the role of reference frame of feedback.

5.1 Randomization and self-selection

Randomization Blocked on school-level, classes were randomized either into the Late-Feedback Treatment or the Early-Feedback Treatment. Within classes students were then randomized into the Control Group, Change Treatment or Level Treatment. Table 7 in Appendix A reports on differences between the class treatments (Late-Feedback Treatment and Early-Feedback Treatment). On average, variables do not differ significantly between the class treatments except with respect to share of participants and teacher experience. Fewer students participate in the study if they are randomized into the Early-Feedback compared to the Late-Feedback Treatment and teachers in the Early-Treatment are more experienced than teachers in the Late-Treatment. The lower share of participants in the Early-Feedback Treatment is surprising as we expected more parents to not give their consent if the intervention is scheduled only a few minutes before the exam. We discuss non-random selection into treatments in detail below. With regard to differences in teachers' experience, we do not think that it affects our treatment effects as observable teacher characteristics such as education or experience do not seem to explain much of the variation in teacher quality [Rivkin et al., 2005]. Tables 8 - 10 in Appendix A present randomization checks for student-level treatments pooled and separately for by class-level treatments. On average, the variables in the Change and Level Treatment do not differ from the Control Group at conventional levels of statistical significance. This indicates that the randomization procedure was successful. However, pupils in the Level-Treatment rate the parents' expectancy that they complete their A-levels higher than pupils in the Change-Treatment. Splitting the sample by class-level treatments shows that this is driven by pupils in the Late.Feedback Treatment. Moreover, pupils who receive the feedback a few days prior to the exam and are in the Change-Feedback Treatment report to have more books at home than pupils in the Control-Treatment. Nevertheless, these differences are small and should not affect our results.

Self-selection into treatments Students in grades 5 and 6 are under-aged and therefore only allowed to participate in the experiment with their parents' consent.[14] Hence, before comparing treatment groups to the control group, concerns about strategic non-participation need to be discussed, as results could lead to biased conclusions if the decision to participate is associated with the outcomes of interest [see Angrist, 1997, Duflo et al., 2007, on non-random samples and selection bias].

Parents in the same class received the same information about the experiment and the same consent forms and students were randomized into student treatments after we obtained parents' consent. Hence, non-random self-selection into the student-level treatments (Control, Change, Level) was not possible. Furthermore, no evidence of non-random self-selection can be found in the randomization table of each class-level treatment (Tables 9 and 10 in Appendix A). In the Early-Treatment, students in the Change-Feedback Treatment seem to have more books than students in the Control Group and in the Late-Treatment, students in the Level-Feedback Treatment are expected to complete their A-levels more often than students in the Change-Feedback Treatment. Nevertheless, this differences are small and only significant on the 10% level.

With respect to class-level treatments (Early, Late) non-random self-selection was possible as parents and students got to know whether the feedback notes will be distributed immediately or a few days prior to the exam. It was not possible to give the same information to parents in the Early- and Late-Treatment as one prerequisite by teachers in the Late-Feedback Treatment to participate was to inform parents that feedback notes will be distributed in the testing hour. Surprisingly, the share of participants turned out to be significantly lower in the Early-Feedback Treatment compared to the Late-Feedback Treatment. We expected the opposite as parents might be concerned about larger negative (emotional) effects of feedback when given shortly before the exam. This could be an indication that parents were not concerned about the timing of the feedback and that the difference in participation rates is just a coincidence. More importantly, there are no significant mean differences between the Control-Group and each of the two treatment groups in measures of past performance (ranks in exam 1 and 2, points in exam 1 and 2, change in rank and share of worseners).

Overall, 157 (30.84%) students did not get their parents' consent to participate in the experiment [68 (28.10%) in the Late-Feedback Treatment and 89 (33.33%) in the Early-Feedback Treatment]. In 16 out of 19 classes, more than 50% of the students within the class participated.[15] These high non-participation rates are

[14] The parents' consent is a necessary legal prerequisite in NRW to conduct scientific studies with under-aged children (see https://www.schulministerium.nrw.de/docs/Recht/Schulrecht/Schulgesetz/Schulgesetz.pdf and http://www.berufsorientierung-nrw.de/cms/upload/BASS_10-45_Nr.2.pdf).

[15] Participation rates ranged from 37.93% - 100% on class-level.

most likely because the intervention was scheduled before a very high-stakes test. However, non-participation rates are also high in low-stakes testing environment. Wagner [2016] studies elementary pupils in the same cities ((Bonn, Cologne and Düsseldorf) in a low-stakes testing environment and finds non-participation rates ranging between 17% - 25%.

We can also check whether pupils included in the study are different from non-participants with respect to past performance. Comparing the grades given by teachers, we find small and insignificant differences exam 1 (Late-Feedback Treatment: 2.61 vs 2.70; Early-Feedback Treatment: 2.90 vs 3.07) and exam 2 (Late-Feedback Treatment: 2.59 vs 2.86; Early-Feedback Treatment: 2.60 vs 2.82)

To summarize, students do not differ across student-level treatments. On class-level treatments, less students participate in the Early-Feedback Treatment. However, students do not differ in past performance measures, also not when compared to non-participants which is why we are not concerned about non-random self-selection

5.2 Data and descriptive statistics

Our data are a mixture of administrative data and questionnaire based data collected on class- and student-level. Importantly, we have very detailed information on students' past performance as we got students' grades in exam 1 and exam 2, students' exact final points in both exam as well as the maximum score possible in the exams. This data can be treated as exogenous in the analysis because they were given to pupils before teachers learned about the experiment and allow to control for heterogeneity in ability. Students are on average 11.64 years old and have 1.61 siblings. 47.55% of the pupils are female and 61.23% speak only German at home. The average grade in exam 1 is 2.75 and 2.59 in exam 2 on a scale from 0.7 to 6, where 0.7 is the highest and 6 is the lowest grade.[16] Figure 1 summarizes the feedback students received by treatment and reveals that students, in part, received an extreme negative or positive feedback. Figures 3 - 4 show the distribution of given feedback pooled over class-level treatments.

Table 1: Descriptive statistics of provided feedback

		Obs.	Mean	Std. Dev	Min.	Max.
Change-Feedback	Early-Feedback	59	0.763	8.052	-21	+21
	Late-Feedback	57	0.842	8.239	-19	+19
Level-Feedback	Early-Feedback	64	13.922	8.407	1	30
	Late-Feedback	60	13.233	8.208	1	30

Note: This table presents descriptive statistics of the feedback given to pupils by class-level and pupil-level treatment.

[16] 0.7=A+; 1.0=A; 1.3=A-; 1,7=B+; 2.0=B; $\cdots$;5.7=F+; 6.0=F.

5.3 Impact of feedback on performance

In the following we will present our results. We will analyze the effect of the timing of feedback (1-3 days before versus immediately before exam) on performance, which was randomized at the class level. Then we will analyze the overall effect of the reference frame of feedback (rank level versus change in rank versus none), which was randomized at the student level. Since we are expecting heterogeneous effects of feedback not only by timing and reference frame but also by whether the content of feedback was positive or negative (positive versus negative change in rank, and high versus low rank), we will, in each case also study the interaction of timing and reference frame with content of feedback. The following tables present results from OLS regressions that include prior performance as linear control variables and student characteristics binary control variables, as well as a constant. In each case the reported standard errors are bootstrapped and clustered at the class level. 2000 repetitions were used for bootstrapping to make sure that the estimates of the standard errors are robust as bootstrapping allows for cluster-robust inference in spite having fewer than 30 clusters Cameron et al. [2008]. Regressions without control variables can be found in the Appendix.

We estimated the following model separately for the class-level treatments:

$$Points_i = \beta_0 + \beta_1\, Feedback_i + \beta_2\, Exam1_i + \beta_2\, Exam2_i + \gamma\, Covariates_i + \delta\, Class_i + \varepsilon_i \tag{1}$$

$Points_i$ are the percentile points in the final math exam (exam 3) of student i, $Exam1_i$ and $Exam2_i$ are the percentile points in exam 1 and exam 2, $Covariates_i$ is a vector of characteristics of student i: students' gender, whether student i has an own room, whether student i speaks a foreign language and the number of siblings, $Class_i$ controls for class fixed effects and ε_i is a stochastic i.i.d. error term.

5.3.1 The role of timing

We first report on differences between classes in the Early-Feedback and Late-Feedback Treatment. Table 2 reports on the effect of receiving the feedback notes 1-3 day prior to the exam pooled by pupil-level treatments. As can bee seen in column 1 students in the Early-Feedback treatment have a 5.6 percentage points higher performance than students in the late treatment. This differences is significant at the 5%-level. Comparing models 2 and 3 we can see that the effect is largely driven by students who decreased their relative performance prior to the exam we study. However, as can be seen in models 4 and 5 there is no evidence that the effect is driven by the students prior level of performance as the coefficient of Early is roughly the same size for those in the better half as compared to those in the worse half.

Table 2: Timing of Feedback - Early vs. Late

Dep. Var: Test Scores	(1) All	(2) If Improved	(3) If Worsened	(4) If Better Half	(5) If Worse Half
Early Feedback	0.056**	0.026	0.092**	0.056*	0.061*
	(0.028)	(0.347)	(0.026)	(0.065)	(0.072)
Points Exam 1	0.242***	0.171	0.416***	0.268**	0.212**
	(0.000)	(0.324)	(0.000)	(0.020)	(0.014)
Points Exam 2	0.390***	0.458**	0.269***	0.378***	0.341***
	(0.000)	(0.012)	(0.006)	(0.000)	(0.008)
Female	−0.022	−0.025	−0.016	−0.024	−0.023
	(0.256)	(0.387)	(0.434)	(0.251)	(0.365)
Controls	Yes	Yes	Yes	Yes	Yes
Observations	322	165	157	170	152
Adjusted R^2	0.371	0.308	0.431	0.217	0.279

Note: This table....Dependent variable: percentage points exam 3. Covariates: percentage points exam 1, percentage points exam 2, gender, own room, foreign name and siblings. The number of clusters is 19, bootstrapped standard errors with 2000 repetitions. p-values in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

The timing of feedback matters and it seems that educators should opt for an early feedback to increase academic performance. However, the significant difference in performance between the two class-level treatments could be caused due to (i) increased learning of students in the Early-Treatment, (ii) peer effects in the Early-Treatment, (iii) an effect of answering the questionnaire[17] or (iv) a negative emotional effect of students in the Late-Treatment.[18] While (i)-(iii) would speak for regularly using an early feedback (iv) would speak against it. In the following we shed light on the underlying mechanism causing differences between the two class-level treatments and show that (ii)-(iv) do not cause the difference.

Whether peer effects and answering the questionnaire have an impact on performance in the exam can be analyzed by comparing the Control Group students of the Early-Feedback Treatment to Control Group students of the Late Treatment. If answering the questionnaire and therefore thinking about self-related beliefs or a change in learning behavior of peers influences students in the Control Group in the Early-Feedback Treatment, they should perform better than students in the Control Group in the Late-Feedback Treatment. However, as can be seen in table 3, this does not seem to be the case. The coefficient is positive but small and insignificant (0.022, p = 0.409). Hence, there are no spillover effects of students receiving a feedback on students receiving no feedback in the Early-Feedback Treatment and no effect of answering the questionnaire.

[17] By answering the questionnaire in the Early-Feedback Treatment, students had to think about their past performance and self-related belief which could have caused their effort in exam preparation or their effort while sitting the exam.

[18] We do not claim that (i)-(iv) are the only candidates which could explain the difference between the two class-level treatments but think that they are the most likely.

Table 3: Timing of Feedback by pupil-level treatments

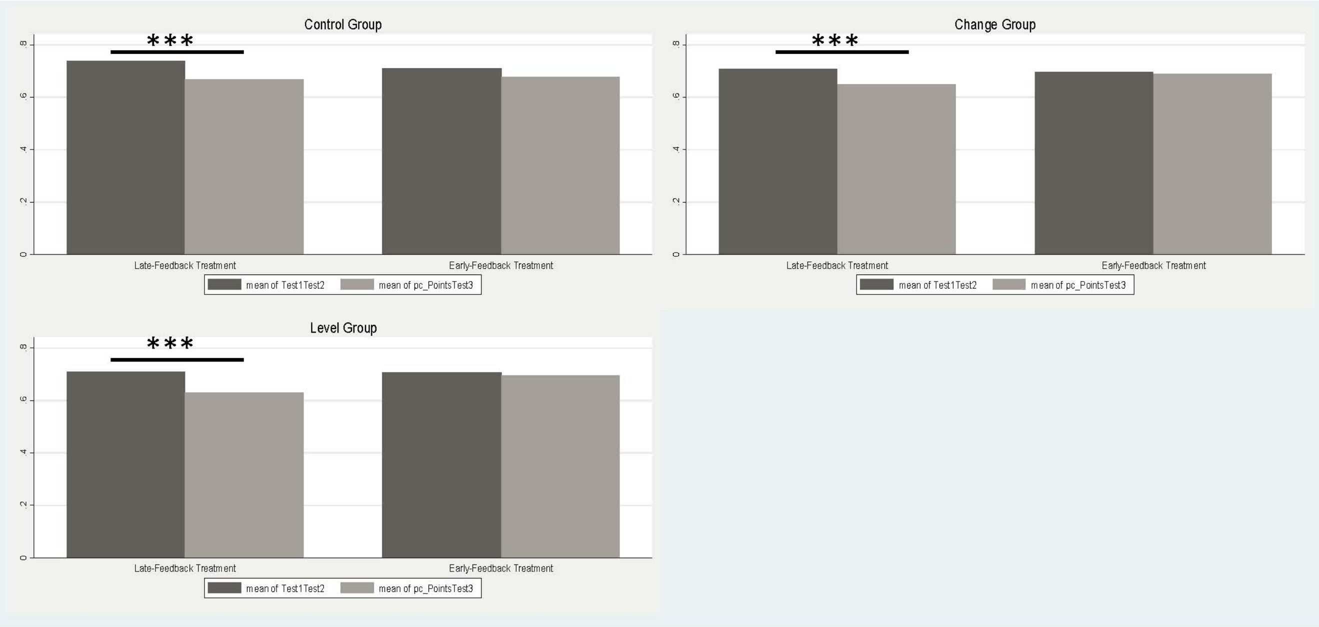
Dep. Var: Test Scores	(1) All	(2) If Improved	(3) If Worsened	(4) If Better Half	(5) If Worse Half
Early Feedback	0.022	0.015	0.038	0.034	0.018
	(0.409)	(0.710)	(0.259)	(0.442)	(0.661)
Change Feedback	−0.005	0.015	−0.017	0.006	−0.017
	(0.713)	(0.678)	(0.605)	(0.832)	(0.683)
Change Feedback × Early Feedback	0.035	−0.028	0.094*	−0.003	0.073
	(0.198)	(0.640)	(0.056)	(0.955)	(0.213)
Level Feedback	−0.025	−0.033	−0.010	−0.043	−0.003
	(0.218)	(0.407)	(0.838)	(0.128)	(0.922)
Level Feedback × Early Feedback	0.061**	0.050	0.066	0.066**	0.057
	(0.016)	(0.374)	(0.209)	(0.034)	(0.363)
Points Exam 1	0.240***	0.164	0.424***	0.261**	0.214**
	(0.000)	(0.323)	(0.000)	(0.019)	(0.022)
Points Exam 2	0.384***	0.461***	0.252***	0.363***	0.327***
	(0.000)	(0.008)	(0.004)	(0.001)	(0.008)
Female	−0.023	−0.026	−0.017	−0.024	−0.028
	(0.230)	(0.425)	(0.452)	(0.275)	(0.375)
Controls	Yes	Yes	Yes	Yes	Yes
Observations	319	163	156	168	151
Adjusted R^2	0.391	0.349	0.476	0.266	0.324

Note: This table....Dependent variable: percentage points exam 3. Covariates: percentage points exam 1, percentage points exam 2, gender, own room, foreign name and siblings. The number of clusters is 19, bootstrapped standard errors with 2000 repetitions. p-values in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Differences between the Late- and Early-Feedback Treatment could be due to an increase in performance of students in the Early-Feedback Treatment or due to a decrease in performance of students in the Late-Feedback Treatment. In order to shed light on the underlying effect, we compare the performance in the last exam to the average past performance (average performance in exam 1 and 2). Figure 2 compares the average past performance in exam 1 and 2 to the performance in exam 3 by class- and school-level treatments. In the Late-Feedback Treatment the performance in the final exam is significantly lower compared to the average performance in exam 1 and 2 in all class-level treatments. In contrast, there is no significant difference between past performance and performance in the final exam for students in the Early-Feedback Treatments. Thus it seems that the difference in performance between the class-level treatments is driven by a decline in performance of students in the Late-Feedback Treatment. However, the performance in the final exam is also lower for students in the Control Group. One reading would be that the final exam is in general harder than the prior exams and that the decline in performance compared to previous exams is a "natural" patern. This in turn leads to the conclusion that the feedback in the Early-Feedback Treatment works positive as it prevents the "natural" decline.

Result 1 *The timing of feedback matters. Students with an early feedback perform better than students with an immediate feedback.*

Figure 2: Past performance vs. performance in exam 3



Note: This figure compare the average past performance (dark gray bars) to the performance in exam 3 (light gray pars) for the Late-Feedback Treatment (left) and the Early-Feedback Treatment (right) separately for each pupil-level treatment.

5.3.2 The role of reference frame of feedback

In the following, we will present results with respect to the reference frame of feedback. We will do so separately for the classes who had the intervention 1-3 days before and the classes who had the intervention immediately before the exam in order to shed light on what is driving the overall better outcomes of students who were treated earlier rather than later. It is particularly important to find out whether the difference in outcomes of early and late treatment classes is driven by early feedback helping students or late feedback harming students, or both types of feedback either helping or harming students but to different degrees. In order to address this question we will compare students who received either level or change feedback with their classmates who did not receive any feedback by including class fixed effects in our model.

Table 4: Class Treatment: Early

	(1)	(2)	(3)	(4)	(5)
Dep. Var: Test Scores	All	If Improved	If Worsened	If Better Half	If Worse Half
Change Feedback	0.038*	0.004	0.083***	0.018	0.076*
	(0.074)	(0.913)	(0.007)	(0.438)	(0.057)
Level Feedback	0.039**	0.026	0.054**	0.034	0.057
	(0.019)	(0.482)	(0.038)	(0.127)	(0.113)
Points Exam 1	0.357***	0.323**	0.468***	0.397***	0.381***
	(0.000)	(0.041)	(0.000)	(0.000)	(0.000)
Points Exam 2	0.296***	0.348***	0.166	0.308	0.051
	(0.001)	(0.000)	(0.162)	(0.435)	(0.564)
Female	0.005	−0.007	0.019	−0.024	0.030
	(0.852)	(0.892)	(0.455)	(0.475)	(0.380)
Controls	Yes	Yes	Yes	Yes	Yes
Class FE	Yes	Yes	Yes	Yes	Yes
Observations	162	88	74	80	82
Adjusted R^2	0.522	0.430	0.618	0.304	0.568

Note: This table . Dependent variable: percentage points exam 3. Covariates: percentage points exam 1, percentage points exam 2, gender, . The number of clusters is . p-values in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Change and level feedback given early Table 4 presents the results with respect to the reference frame of feedback for classes who were treated 1-3 days before the exam. As can be seen in model 1 both types of feedback seem to help students as compared to the control group within their class who did not receive any feedback. Students who received change and students who received level feedback both have 3.8 percentile points higher outcomes than students in the control group, although the effects are only significant at the 10%-level and the 5%-level, respectively. As can be seen in models 2 and 3 the effect is largely driven by giving feedback to students who decreased their relative performance prior to the last exam. Telling students who decreased their relative performance by how much their relative performance decreased increases their performance in the final test by 8.3 percentile points as compared to their classmates who got worse but received no feedback. This effect is significant at the 1%-level. Students who got worse and who received level feedback have a 5.4 percentile points better than students who received no feedback. This effect is significant at the 5%-level. F-tests show that the coefficients of the change feedback and the level feedback are not significantly different from each other. There is weak evidence that giving change feedback (positive or negative) to students who had a below median performance in the last test improves their performance in the following test (see column 5). Models 4 and 5 do not suggest that there is a significant interaction of level feedback with prior level of performance.

Table 5: Class Treatment: Late

	(1)	(2)	(3)	(4)	(5)
Dep. Var: Test Scores	All	If Improved	If Worsened	If Better Half	If Worse Half
Change Feedback	−0.001	0.023	−0.029	0.011	0.011
	(0.954)	(0.581)	(0.276)	(0.663)	(0.756)
Level Feedback	−0.022	−0.0071	−0.023	−0.035	−0.020
	(0.253)	(0.819)	(0.614)	(0.217)	(0.626)
Points Exam 1	0.123	0.092	0.382^{***}	0.179	0.008
	(0.409)	(0.709)	(0.000)	(0.220)	(0.896)
Points Exam 2	0.437^{***}	0.434	0.256^{*}	0.332^{***}	0.475^{***}
	(0.000)	(0.194)	(0.094)	(0.003)	(0.000)
Female	−0.041	−0.046	−0.021	−0.022	−0.065
	(0.220)	(0.253)	(0.450)	(0.579)	(0.112)
Controls	Yes	Yes	Yes	Yes	Yes
Class FE	Yes	Yes	Yes	Yes	Yes
Observations	160	77	83	90	70
Adjusted R^2	0.363	0.208	0.456	0.156	0.307

Note: This table . Dependent variable: percentage points exam 3. Covariates: percentage points exam 1, percentage points exam 2, gender, . The number of clusters is . p-values in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Change and level feedback given late Table 5 presents the results with respect to the reference frame of feedback for classes that were treated immediately before the exam. Overall, we can see that none of the coefficients of the treatment dummies in any of the models are significant. Furthermore, the overall effect of the change feedback is very close to zero (column 1) but there seems to be heterogeneity in effects. The coefficient of the change feedback treatment dummy has a positive sign for students who improved (column 2) and a negative sign for students who got worse (column 3), although none of them are significant, which as a very weak indication that the effect of change feedback given immediately before the exam depends on whether the feedback is positive or negative. There is no evidence that the effect of level feedback depends on the content, as the coefficient of the level feedback dummy is negative and of similar magnitude both for the better (column 4) and in the worse half of students (column 5).

Overall, our results indicate that both change and level feedback, when given 1-3 days before the exam have a positive effect on subsequent performance and is particularly beneficial for students who recently decreased their performance. We do not find significant effects of feedback given immediately before the exam. However, the signs of the coefficients are a very weak indication that level feedback overall as well as negative change feedback have a negative effect on performance when administered immediately before an exam.

Result 2 *The effect of feedback on subsequent performance depends on the timing of feedback and prior changes in performance but not on whether feedback is given in terms of changes or levels of performance.*

6 Mechanisms

Which students react to the feedback, and in what ways? Do girls react differently than boys? Does competitiveness, math confidence and self-esteem matter for how a child reacts to feedback? We will address these questions in the following. We can study the interaction of our treatment with gender and competitiveness both for the early and the late treatment as these variables were included in questionnaires both for the early and the late treatment. The questionnaire in the early treatment contained additional scales, such as on math confidence and self-esteem, which were not included in the questionnaire of the late treatment due to time constraints as this questionnaire was to be answered during the same lesson the exam was written. In this section we will try to shed light on the behavioral mechanisms driving our results. First, we will try to understand whether our results are driven by a certain subgroup. In particular, we will be discussing the role of gender and character traits (competitiveness, confidence, and self-esteem) in explaining our results. Second, we will look at whether the effect of early feedback on outcomes can be explained by a change in beliefs about the effectiveness of learning effort.

We find that the overall positive effect of both change and level feedback in the early treatment is driven by the response of boys (see Table 11 in the Appendix). Boys have 5.9 and 7.4 percentage points better results in the change and level treatments, respectively, than in the control group. At the same time, there is no significant difference for girls in any of the two treatment groups and the control group. The coefficients of the treatment dummies and the interaction term of treatment and the female indicator each add up to an almost perfect zero effect. Looking at improvers and worseners separately, there is a positive effect of level feedback on boys who improved but no effect of any type of feedback on girls who improved, as F-tests show that the combined coefficients of the treatment dummies and the female indicators are not significantly different from zero. We also find that both boys and girls respond positively to feedback about negative changes, as the coefficient of the interaction term of change feedback and female is very small an insignificant. Furthermore, there is a positive effect of both change and level feedback on boys who are in the worse half. F-tests show that the effect on girls is not significantly different from zero.

Splitting the sample at the median value of the competitiveness, confidence, and self-esteem measures, we find that competitiveness and confidence do not interact with our feedback intervention. However, we find the overall positive effects of level and change feedback given 1-3 days before an exam is driven by students who report low self-esteem.

7 Conclusion

We have tested an inexpensive and easy to implement feedback intervention in secondary schools in Germany. We varied the timing and reference frame of relative performance feedback to analyze the causal effect on performance in a high-stakes exam. With respect to timing, we compare students who received a feedback note either 1-3 days before the last math exam of the semester to students receiving the feedback in the same lesson immediately before the exam started. Concerning reference frame of feedback, students in the Control Group got “good luck” wishes while students in the treatment group got either a level feedback—the absolute rank in the preceding exam—or a change feedback—the change in ranks between the two preceding exams.

We find that the timing of rank feedback is essential. It is harmful immediately before the test but useful if given 1-3 days in advance. Then it is especially useful for those who got worse. Moreover, the timing of feedback about the change in performance is also essential. It is not found to be effective if given immediately before the test. But explicitly telling people, who got worse, that they got worse 1-3 days in advance improves their performance. Finally, we do not find evidence that any of the two feedback types affect pupils’ self-related beliefs such as locus of control, self-esteem, academic and math efficacy or perceived effectiveness of learning

Our results give interesting insights into how relative feedback works in educational settings and has important implications for educational policy makers and teachers how to best design feedback. Moreover, our findings indicate that teachers should not only give positive feedback—sweet treats—but also consider to openly communicate whenever pupils’ performance worsened—bitter pills.

References

Joshua Angrist. Conditional Independence in Sample Selection Models. *Economics Letters*, 54(2):103–112, 1997.

Nava Ashraf, Oriana Bandiera, and Scott S. Lee. Awards unbundled: Evidence from a natural field experiment. *Journal of Economic Behavior & Organization*, 100:44 – 63, 2014. ISSN 0167-2681. doi: http://dx.doi.org/10.1016/j.jebo.2014.01.001. URL http://www.sciencedirect.com/science/article/pii/S0167268114000079.

Ghazala Azmat and Nagore Iriberri. The importance of relative performance feedback information: Evidence from a natural experiment using high school students. *Journal of Public Economics*, 94(7-8):435 – 452, 2010.

Ghazala Azmat and Nagore Iriberri. The Provision of Relative Performance Feedback: An Analysis of Performance and Satisfaction. *Journal of Economics & Management Strategy*, 25(1):77–110, 2016.

Ghazala Azmat, Manuel Bagues, Antonio Cabrales, and Nagore Iriberri. What you don't know... Can't hurt you? A field experiment on relative performance feedback in higher education. Discussion Paper DP11201, Centre for Economic Policy Research, March 2016.

Oriana Bandiera, Valentino Larcinese, and Imran Rasul. Blissful ignorance? a natural experiment on the effect of feedback on students' performance. *Labour Economics*, 34:13 – 25, 2015. ISSN 0927-5371. doi: http://dx.doi.org/10.1016/j.labeco.2015.02.002. URL http://www.sciencedirect.com/science/article/pii/S092753711500010X. European Association of Labour Economists 26th Annual Conference.

Iwan Barankay. Rank incentives - evidence from a randomized workplace experiment. *unpublished working paper*, 2012. URL https://mgmt.wharton.upenn.edu/files/?whdmsaction=public:main.file&fileID=4357.

Eric Bettinger. Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores. *Review of Economics and Statistics*, 94(3):686–698, 2012.

Christiane Bradler, Robert Dur, Susanne Neckermann, and Arjan Non. Employee Recognition and Performance: A Field Experiment. *Management Science*, accepted, 2016.

Colin Cameron, Jonah Gelbach, and Douglas Miller. Bootstrap-Based Improvements for Inference with Clustered Errors. *The Review of Economics and Statistics*, 90(3):414–427, 2008.

Gary Charness and Matthew Rabin. Understanding social preferences with simple tests*. *The Quarterly Journal of Economics*, 117(3):817, 2002. doi: 10.1162/003355302760193904. URL +http://dx.doi.org/10.1162/003355302760193904.

Gary Charness, David Masclet, and Marie Claire Villeval. The dark side of competition for status. *Management Science*, 60(1):38–55, 2014. doi: 10.1287/mnsc.2013.1747. URL http://dx.doi.org/10.1287/mnsc.2013.1747.

Flavio Cunha and James Heckman. The Technology of Skill Formation. *American Economic Review*, 97(2): 31–47, 2007.

David Cutler and Adriana Lleras-Muney. Education and Health: Evaluating Theories and Evidence. Working Paper 12352, National Bureau of Economic Research, July 2006.

Mette Trier Damgaard and Helena Skyt Nielsen. The use of nudges and other behavioural approaches in education. EENEE Analytical Report 29, Prepared for the European Commission, February 2017.

Emmanuel Dechenaux, Dan Kovenock, and Roman M. Sheremeta. A survey of experimental research on contests, all-pay auctions and tournaments. *Experimental Economics*, 18(4):609–669, 2015. ISSN 1573-6938. doi: 10.1007/s10683-014-9421-0. URL http://dx.doi.org/10.1007/s10683-014-9421-0.

Esther Duflo, Rachel Glennerster, and Michael Kremer. Chapter 61 - Using Randomization in Development Economics Research: A Toolkit. In Paul Schultz and John Strauss, editors, *Handbook of Development Economics*, volume 4, pages 3895–3962. North-Holland, Elsevier, 2007.

Tor Eriksson, Anders Poulsen, and Marie Claire Villeval. Feedback and incentives: Experimental evidence. *Labour Economics*, 16(6):679–688, 2009. ISSN 0927-5371. doi: https://doi.org/10.1016/j.labeco.2009.08.006. URL http://www.sciencedirect.com/science/article/pii/S0927537109000980.

Armin Falk and Andrea Ichino. Clean evidence on peer effects. *Journal of labor economics*, 24(1):39–57, 2006.

Roland Fryer. Teacher Incentives and Student Achievement: Evidence from New York City Public Schools. *Journal of Labor Economics*, 31:373–427, 2013.

Roland Fryer, Steven Levitt, John List, and Sally Sadoff. Enhancing the Efficacy of Teacher Incentives through Loss Aversion: A Field Experiment. Working Paper 18237, National Bureau of Economic Research, July 2012.

David Gill, Victoria Prowse, Zdenka Kissova, and Jaesun Lee. First-Place Loving and Last-Place Loathing: How Rank in the Distribution of Performance Affects Effort Provision. Discussion Paper 783, Oxford Department of Economics, March 2016.

Oliver Gürtler and Christine Harbring. Feedback in tournaments under commitment problems: Experimental evidence. *Journal of Economics & Management Strategy*, 19(3):771–810, 2010. ISSN 1530-9134. doi: 10.1111/j.1530-9134.2010.00269.x. URL http://dx.doi.org/10.1111/j.1530-9134.2010.00269.x.

Lynn Hannan, Ranjani Krishnan, and Andrew Newman. The effects of disseminating relative performance feedback in tournament and individual performance compensation plans. *The Accounting Review*, 83(4): 893–913, 2008. doi: 10.2308/accr.2008.83.4.893. URL http://dx.doi.org/10.2308/accr.2008.83.4.893.

Lynn Hannan, Gregory McPhee, Andrew Newman, and Ivo Tafkov. The effect of relative performance information on performance and effort allocation in a multi-task environment. *The Accounting Review*, 88 (2):553–575, 2013. doi: 10.2308/accr-50312. URL http://dx.doi.org/10.2308/accr-50312.

Eric Hanushek and Steven Rivkin. Teacher Quality. In Eric Hanushek and Finis Welch, editors, *Handbook of the Economics of Education*, volume 2, pages 1051–1078. Elsevier, 2006.

Eric Hanushek, Guido Schwerdt, Simon Wiederhold, and Ludger Wößmann. Returns to Skills around the World: Evidence from {PIAAC}. *European Economic Review*, 73:103–130, 2015.

Nina Jalava, Juanna Schroter Joensen, and Elin Pellas. Grades and rank: Impacts of non-financial incentives on test performance. *Journal of Economic Behavior & Organization*, 115:161–196, 2015. ISSN 0167-2681. doi: http://dx.doi.org/10.1016/j.jebo.2014.12.004. URL http://www.sciencedirect.com/science/article/pii/S0167268114003163.

Avraham N. Kluger and Angelo DeNisi. The Effects of Feedback Interventions on Performance: A historical review, a meta-analysis, and a preliminary Feedback Intervention Theory. *Psychological Bulletin*, 119 (2):254–284, 1996. URL http://mario.gsia.cmu.edu/micro_2007/readings/feedback_effects_meta_analysis.pdf.

Avraham N. Kluger and Angelo DeNisi. Feedback Interventions: Toward the Understanding of a Double-Edged Sword. *Current Directions in Psychological Science*, 7(3):pp. 67–72, 1998. ISSN 09637214. URL http://www.jstor.org/stable/20182507.

Camelia M. Kuhnen and Agnieszka Tymula. Feedback, self-esteem, and performance in organizations. *Management Science*, 58(1):94–113, 2012. doi: 10.1287/mnsc.1110.1379. URL http://dx.doi.org/10.1287/mnsc.1110.1379.

Ilyana Kuziemko, Ryan W. Buell, Taly Reich, and Michael I. Norton. "last-place aversion": Evidence and redistributive implications. *The Quarterly Journal of Economics*, 129(1):105, 2014. doi: 10.1093/qje/qjt035. URL +http://dx.doi.org/10.1093/qje/qjt035.

Steven Levitt, John List, Susanne Neckermann, and Sally Sadoff. The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance. *American Economic Journal: Economic Policy*, 8(4):183–219, 2016a.

Steven Levitt, John List, and Sally Sadoff. The effect of performance-based incentives on educational achievement: Evidence from a randomized experiment. NBER Working Paper 22107, National Bureau of Economic Research, march 2016b.

Alexandre Mas and Enrico Moretti. Peers at work. *The American Economic Review*, 99(1):112–145, 2009.

Kevin Milligan, Enrico Moretti, and Philip Oreopoulos. Does education improve citizenship? evidence from the united states and the united kingdom. *Journal of Public Economics*, 88(9 - 10):1667 – 1695, 2004.

Philip Oreopoulos. Do Dropouts Drop out Too Soon? Wealth, Health and Happiness from Compulsory Schooling. *Journal of Public Economics*, 91(11):2213–2229, 2007.

Philip Oreopoulos and Kjell Salvanes. Priceless: The nonpecuniary benefits of schooling. *The Journal of Economic Perspectives*, 25(1):159–184, 2011.

Eleanor O'Rourke, Kyla Haimovitz, Christy Ballweber, Carol Dweck, and Zoran Popović. Brain points: A growth mindset incentive structure boosts persistence in an educational game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3339–3348. ACM, 2014.

David Paunesku, Gregory Walton, Carissa Romero, Eric Smith, David Yeager, and Carol Dweck. Mind-Set Interventions Are a Scalable Treatment for Academic Underachievement. *Psychological Science*, 26(6): 784–793, 2015.

Steven Rivkin, Eric Hanushek, and John Kain. Teachers, Schools, and Academic Achievement. *Econometrica*, 73(2):417–458, 2005.

Julian Rotter. Generalized expectancies for internal versus external control of reinforcement. *Psychological monographs: General and applied*, 80(1):1, 1966.

Adam Smith. The theory of moral sentiments. *London: Printed for A. Millar, and A. Kincaid and J. Bell.*, 1759.

Ivo D. Tafkov. Private and public relative performance information under different compensation contracts. *The Accounting Review*, 88(1):327–350, 2013. doi: 10.2308/accr-50292. URL http://dx.doi.org/10.2308/accr-50292.

Richard Thaler, Cass Sunstein, and John Balz. Chapter 25 - Choice Architecture. In Eldar Shafir, editor, *The Behavioral Foundations of Public Policy*, pages 428–439. Princeton University Press, 2013.

Anh Tran and Richard Zeckhauser. Rank as an inherent incentive: Evidence from a field experiment. *Journal of Public Economics*, 96(9-10):645–650, 2012. ISSN 0047-2727. doi: http://dx.doi.org/10.1016/j.jpubeco.2012.05.004. URL http://www.sciencedirect.com/science/article/pii/S0047272712000436.

Valentin Wagner. Seeking Risk or Answering Smart? Framing in Elementary Schools. Discussion Paper 227, Düsseldorf Institute for Competition Economics (DICE), October 2016.

Valentin Wagner and Gerhard Riener. Peers or Parents? On Non-Monetary Incentives in Schools. DICE Discussion Papers 203, Heinrich-Heine-Universität Düsseldorf, Düsseldorf Institute for Competition Economics (DICE), 2015.

Appendix

A Tables

A.1 Balance and randomization checks

Table 6: Treatment Observations

		Class Level Randomization		
		Late-Feedback Treatment	Early-Feedback Treatment	*Total Observations*
Pupil Level Randomization	Change Treatment	57	59	116
	Level Treatment	61	64	125
	Control Treatment	56	55	111
	Total Observations	174	178	*352*

Note: This table summarizes the number of participants by treatment groups. In total, 352 children in 19 classes in 7 schools received parents' consent and participated.

Table 7: Randomization Check Class Treatments

	(1) Late-Feedback Treatment	(2) Early-Feedback Treatment	(3) Overall	(4) (1) vs. (2), p-value
Female Teacher	0.793	0.781	0.787	0.781
	(0.031)	(0.031)	(0.022)	
Class Size	27.782	27.242	27.509	0.123
	(0.244)	(0.250)	(0.175)	
Age	23.667	24.708	24.193	0.363
	(0.816)	(0.802)	(0.572)	
Points Exam1	0.712	0.681	0.696	0.105
	(0.014)	(0.014)	(0.010)	
Points Exam2	0.719	0.730	0.725	0.554
	(0.014)	(0.013)	(0.009)	
Rank Exam1	0.495	0.490	0.493	0.889
	(0.022)	(0.021)	(0.015)	
Rank Exam2	0.467	0.493	0.481	0.399
	(0.021)	(0.022)	(0.015)	
Change in Rank	0.523	−0.028	0.243	0.505
	(0.592)	(0.577)	(0.413)	
Share Worsen	0.506	0.455	0.480	0.343
	(0.038)	(0.037)	(0.027)	
Share Participants	0.775	0.703	0.739	0.000
	(0.015)	(0.012)	(0.010)	
Female Pupil	0.480	0.449	0.464	0.570
	(0.038)	(0.037)	(0.027)	
Single Room	0.655	0.596	0.625	0.370
	(0.046)	(0.048)	(0.033)	
Internet	1.115	1.022	1.068	0.366
	(0.072)	(0.073)	(0.051)	
A-Level	2.034	2.056	2.045	0.879
	(0.103)	(0.099)	(0.071)	
Car	1.333	1.303	1.318	0.785
	(0.078)	(0.078)	(0.055)	
Siblings	1.299	1.489	1.395	0.165
	(0.094)	(0.099)	(0.068)	
Teacher Exp.	9.902	12.833	11.513	0.008
	(0.647)	(0.831)	(0.548)	
Books at Home	1.983	2.140	2.063	0.314
	(0.110)	(0.111)	(0.078)	
N	174	178	352	
Proportion	0.494	0.506	1.000	

Note: This table presents randomization checks between the Late-Feedback and Early-Feedback Treatment. Standard errors in parentheses.

Table 8: Randomization Check Pupil Treatment- Overall

	(1) Control	(2) Change	(3) Level	(4) Overall	(5) (1) vs. (2), p-value	(6) (1) vs. (3), p-value	(7) (2) vs. (3), p-value
Female Teacher	0.782	0.784	0.790	0.786	0.961	0.875	0.912
	(0.040)	(0.038)	(0.037)	(0.022)			
Class Size	27.518	27.595	27.403	27.503	0.860	0.792	0.655
	(0.311)	(0.301)	(0.304)	(0.176)			
Age	23.230	22.745	22.857	22.937	0.654	0.738	0.917
	(0.789)	(0.739)	(0.781)	(0.444)			
Points Exam1	0.718	0.685	0.696	0.699	0.176	0.333	0.617
	(0.017)	(0.018)	(0.015)	(0.010)			
Points Exam2	0.731	0.722	0.722	0.725	0.676	0.668	0.996
	(0.016)	(0.016)	(0.016)	(0.009)			
Rank Exam1	0.455	0.506	0.505	0.490	0.189	0.173	0.968
	(0.027)	(0.028)	(0.024)	(0.015)			
Rank Exam2	0.470	0.479	0.492	0.481	0.811	0.550	0.714
	(0.028)	(0.026)	(0.026)	(0.015)			
Change in Rank	−0.491	0.802	0.371	0.243	0.213	0.383	0.672
	(0.706)	(0.753)	(0.686)	(0.413)			
Share Worsen	0.491	0.500	0.460	0.483	0.892	0.635	0.534
	(0.048)	(0.047)	(0.045)	(0.027)			
Share Participants	0.744	0.736	0.733	0.737	0.719	0.641	0.920
	(0.017)	(0.017)	(0.016)	(0.010)			
Female Pupil	0.418	0.483	0.488	0.464	0.332	0.289	0.938
	(0.047)	(0.047)	(0.045)	(0.027)			
Single Room	0.755	0.785	0.754	0.765	0.607	0.993	0.591
	(0.043)	(0.040)	(0.040)	(0.024)			
Internet	1.168	1.286	1.325	1.263	0.256	0.131	0.709
	(0.072)	(0.074)	(0.073)	(0.042)			
A-level	2.427	2.388	2.593	2.472	0.733	0.117	0.034
	(0.087)	(0.073)	(0.062)	(0.043)			
Car	1.451	1.570	1.586	1.538	0.270	0.202	0.885
	(0.075)	(0.078)	(0.074)	(0.044)			
Siblings	1.343	1.267	1.368	1.327	0.435	0.794	0.276
	(0.072)	(0.067)	(0.065)	(0.039)			
Teacher Exp.	11.349	11.517	11.567	11.482	0.902	0.871	0.970
	(0.968)	(0.965)	(0.930)	(0.549)			
Books at Home	2.196	2.434	2.381	2.340	0.149	0.247	0.748
	(0.110)	(0.121)	(0.114)	(0.067)			
N	110	116	124	350			
Proportion	0.314	0.331	0.354	1.000			

Note: This table presents randomization checks for the pooled Late and Early-Feedback Treatment. Standard errors in parentheses.

Table 9: Randomization Check Pupil Treatment- Late-Feedback Treatment

	(1) Control	(2) Change	(3) Level	(4) Overall	(5) (1) vs. (2), p-value	(6) (1) vs. (3), p-value	(7) (2) vs. (3), p-value
Female Teacher	0.782	0.789	0.800	0.791	0.922	0.813	0.889
	(0.056)	(0.054)	(0.052)	(0.031)			
Class Size	27.782	27.877	27.667	27.773	0.874	0.852	0.730
	(0.429)	(0.421)	(0.437)	(0.247)			
Age	22.667	22.075	22.429	22.382	0.712	0.885	0.823
	(1.174)	(1.086)	(1.136)	(0.650)			
Points Exam1	0.745	0.708	0.703	0.718	0.264	0.179	0.871
	(0.022)	(0.024)	(0.022)	(0.013)			
Points Exam2	0.730	0.712	0.717	0.719	0.581	0.681	0.881
	(0.024)	(0.024)	(0.023)	(0.014)			
Rank Exam1	0.438	0.502	0.522	0.489	0.253	0.105	0.706
	(0.039)	(0.040)	(0.034)	(0.022)			
Rank Exam2	0.457	0.470	0.475	0.467	0.800	0.728	0.924
	(0.038)	(0.036)	(0.036)	(0.021)			
Change in Rank	−0.600	0.842	1.250	0.523	0.342	0.190	0.777
	(1.044)	(1.091)	(0.943)	(0.592)			
Share Worsen	0.527	0.544	0.467	0.512	0.862	0.520	0.408
	(0.068)	(0.067)	(0.065)	(0.038)			
Share Participants	0.778	0.772	0.770	0.773	0.861	0.812	0.953
	(0.026)	(0.026)	(0.025)	(0.015)			
Female Pupil	0.418	0.544	0.475	0.480	0.186	0.549	0.460
	(0.067)	(0.067)	(0.066)	(0.038)			
Single Room	0.745	0.811	0.804	0.787	0.421	0.474	0.919
	(0.062)	(0.054)	(0.054)	(0.032)			
Internet	1.235	1.255	1.411	1.304	0.898	0.220	0.278
	(0.107)	(0.108)	(0.095)	(0.059)			
A-level	2.511	2.320	2.604	2.480	0.251	0.518	0.059
	(0.113)	(0.119)	(0.091)	(0.063)			
Car	1.431	1.491	1.655	1.528	0.694	0.168	0.309
	(0.106)	(0.106)	(0.120)	(0.064)			
Siblings	1.220	1.245	1.268	1.245	0.866	0.742	0.874
	(0.108)	(0.104)	(0.097)	(0.059)			
Teacher Exp.	9.795	9.725	9.930	9.820	0.966	0.933	0.897
	(1.159)	(1.132)	(1.098)	(0.647)			
Books at Home	2.160	2.189	2.382	2.247	0.900	0.361	0.409
	(0.167)	(0.155)	(0.173)	(0.095)			
N	55	57	60	172			
Proportion	0.320	0.331	0.349	1.000			

Note: This table presents randomization checks for students in the Late-Feedback Treatment. Standard errors in parentheses.

Table 10: Randomization Check Pupil Treatment- Early-Feedback Treatment

	(1) Control	(2) Change	(3) Level	(4) Overall	(5) (1) vs. (2), p-value	(6) (1) vs. (3), p-value	(7) (2) vs. (3), p-value
Female Teacher	0.782	0.780	0.781	0.781	0.978	0.994	0.983
	(0.056)	(0.054)	(0.052)	(0.031)			
Class Size	27.255	27.322	27.156	27.242	0.914	0.874	0.784
	(0.452)	(0.429)	(0.424)	(0.250)			
Age	23.750	23.415	23.286	23.478	0.820	0.761	0.930
	(1.069)	(1.005)	(1.080)	(0.604)			
Points Exam1	0.691	0.661	0.690	0.681	0.422	0.967	0.407
	(0.025)	(0.027)	(0.021)	(0.014)			
Points Exam2	0.733	0.732	0.727	0.730	0.976	0.845	0.866
	(0.023)	(0.022)	(0.021)	(0.013)			
Rank Exam1	0.472	0.510	0.488	0.490	0.487	0.751	0.678
	(0.038)	(0.038)	(0.035)	(0.021)			
Rank Exam2	0.482	0.487	0.508	0.493	0.934	0.637	0.687
	(0.041)	(0.038)	(0.037)	(0.022)			
Change in Rank	−0.382	0.763	−0.453	−0.028	0.424	0.959	0.400
	(0.959)	(1.048)	(0.988)	(0.577)			
Share Worsen	0.455	0.458	0.453	0.455	0.974	0.988	0.960
	(0.068)	(0.065)	(0.063)	(0.037)			
Share Participants	0.710	0.701	0.699	0.703	0.751	0.706	0.959
	(0.022)	(0.021)	(0.020)	(0.012)			
Female Pupil	0.418	0.424	0.500	0.449	0.953	0.376	0.401
	(0.067)	(0.065)	(0.063)	(0.037)			
Single Room	0.765	0.759	0.707	0.742	0.948	0.500	0.536
	(0.060)	(0.059)	(0.060)	(0.034)			
Internet	1.100	1.315	1.241	1.222	0.129	0.345	0.628
	(0.096)	(0.102)	(0.111)	(0.060)			
A-level	2.347	2.453	2.582	2.465	0.500	0.130	0.292
	(0.132)	(0.088)	(0.085)	(0.059)			
Car	1.471	1.648	1.518	1.547	0.255	0.731	0.363
	(0.106)	(0.113)	(0.088)	(0.059)			
Siblings	1.462	1.288	1.466	1.407	0.170	0.975	0.145
	(0.093)	(0.084)	(0.086)	(0.051)			
Teacher Exp.	12.638	12.980	12.870	12.833	0.870	0.910	0.957
	(1.471)	(1.466)	(1.409)	(0.831)			
Books at Home	2.231	2.679	2.379	2.429	0.057	0.481	0.205
	(0.144)	(0.182)	(0.151)	(0.093)			
N	55	59	64	178			
Proportion	0.309	0.331	0.360	1.000			

Note: This table presents randomization checks for students in the Early-Feedback Treatment. Standard errors in parentheses.

A.2 Sub-group analyses

Table 11: Class Treatment: Early gender

Dep. Var: Test Scores	(1) All	(2) If Improved	(3) If Worsened	(4) If Better Half	(5) If Worse Half
Change Feedback	0.059***	0.051	0.087*	0.042	0.099**
	(0.000)	(0.328)	(0.079)	(0.329)	(0.024)
Change × Female	−0.050*	−0.121**	−0.008	−0.061	−0.051
	(0.079)	(0.035)	(0.930)	(0.313)	(0.657)
Level Feedback	0.074***	0.096***	0.066	0.075*	0.089***
	(0.000)	(0.000)	(0.193)	(0.053)	(0.000)
Level × Female	−0.073**	−0.162***	−0.022	−0.094	−0.064
	(0.017)	(0.002)	(0.771)	(0.126)	(0.402)
Points Exam 1	0.363***	0.328***	0.477***	0.381***	0.393***
	(0.000)	(0.000)	(0.005)	(0.000)	(0.000)
Points Exam 2	0.292**	0.337***	0.154	0.338	0.045
	(0.013)	(0.000)	(0.257)	(0.378)	(0.612)
Female	0.049	0.094*	0.029	0.031	0.071
	(0.230)	(0.055)	(0.588)	(0.487)	(0.449)
Controls	Yes	Yes	Yes	Yes	Yes
Class FE	Yes	Yes	Yes	Yes	Yes
Observations	162	88	74	80	82
Adjusted R^2	0.523	0.449	0.604	0.294	0.562

Note: This table . Dependent variable: percentage points exam 3. Covariates: percentage points exam 1, percentage points exam 2, gender, . The number of clusters is . p-values in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table 12: Class Treatment: Late gender

	(1)	(2)	(3)	(4)	(5)
Dep. Var: Test Scores	All	If Improved	If Worsened	If Better Half	If Worse Half
Change Feedback	−0.012	−0.003	−0.030	0.018	0.012
	(0.740)	(0.951)	(0.450)	(0.854)	(0.848)
Change × Female	0.020	0.073	0.006	−0.020	−0.003
	(0.731)	(0.256)	(0.987)	(0.893)	(0.972)
Level Feedback	−0.014	−0.039	0.018	−0.015	−0.022
	(0.598)	(0.225)	(0.843)	(0.744)	(0.594)
Level × Female	−0.018	0.086	−0.073	−0.045	0.003
	(0.720)	(0.246)	(0.408)	(0.544)	(0.983)
Points Exam 1	0.121	0.080	0.412***	0.176	0.007
	(0.453)	(0.863)	(0.000)	(0.218)	(0.958)
Points Exam 2	0.434***	0.438	0.227	0.340***	0.476***
	(0.000)	(0.212)	(0.152)	(0.000)	(0.000)
Female	−0.041	−0.104**	0.001	0.002	−0.065
	(0.252)	(0.031)	(0.914)	(0.935)	(0.311)
Controls	Yes	Yes	Yes	Yes	Yes
Class FE	Yes	Yes	Yes	Yes	Yes
Observations	160	77	83	90	70
Adjusted R^2	0.356	0.191	0.451	0.136	0.280

Note: This table . Dependent variable: percentage points exam 3. Covariates: percentage points exam 1, percentage points exam 2, gender, . The number of clusters is . p-values in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table 13: Class Treatment: Early comp

Dep. Var: Test Scores	(1) All	(2) If Improved	(3) If Worsened	(4) If Better Half	(5) If Worse Half
Change Feedback	0.038	−0.073	0.129*	0.120	0.040
	(0.381)	(0.345)	(0.059)	(0.133)	(0.765)
Change × High Comp.	−0.004	0.101*	−0.092	−0.134	0.079
	(0.943)	(0.070)	(0.253)	(0.161)	(0.461)
Level Feedback	0.047	−0.017	0.124	0.113**	0.074
	(0.428)	(0.803)	(0.156)	(0.022)	(0.677)
Level × High Comp.	−0.020	0.040	−0.119	−0.126**	−0.039
	(0.790)	(0.631)	(0.275)	(0.028)	(0.850)
High Comp.	−0.031	−0.094**	0.047	0.025	−0.016
	(0.547)	(0.022)	(0.473)	(0.580)	(0.867)
Points Exam 1	0.334***	0.290	0.528***	0.339**	0.391***
	(0.000)	(0.142)	(0.000)	(0.035)	(0.000)
Points Exam 2	0.316***	0.373**	0.120	0.315	−0.001
	(0.000)	(0.000)	(0.274)	(0.285)	(0.999)
Female	−0.003	−0.017	0.013	−0.040	0.028
	(0.856)	(0.749)	(0.520)	(0.270)	(0.286)
Controls	Yes	Yes	Yes	Yes	Yes
Class FE	Yes	Yes	Yes	Yes	Yes
Observations	162	88	74	80	82
Adjusted R^2	0.524	0.430	0.630	0.324	0.570

Note: This table . Dependent variable: percentage points exam 3. Covariates: percentage points exam 1, percentage points exam 2, gender, . The number of clusters is . p-values in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table 14: Class Treatment: Late comp

Dep. Var: Test Scores	(1) All	(2) If Improved	(3) If Worsened	(4) If Better Half	(5) If Worse Half
Change Feedback	0.036	0.106	−0.016	0.058	0.017
	(0.323)	(0.282)	(0.650)	(0.343)	(0.644)
Change × High Comp.	−0.063	−0.130	−0.031	−0.082	−0.013
	(0.257)	(0.341)	(0.777)	(0.451)	(0.831)
Level Feedback	−0.031	−0.061	0.052	−0.036	−0.060
	(0.479)	(0.438)	(0.446)	(0.639)	(0.147)
Level × High Comp.	0.011	0.096	−0.125	−0.002	0.061
	(0.861)	(0.314)	(0.185)	(0.995)	(0.528)
High Comp.	0.010	0.021	0.042	0.011	−0.014
	(0.754)	(0.649)	(0.497)	(0.878)	(0.733)
Points Exam 1	0.112	0.066	0.320***	0.166	−0.007
	(0.395)	(0.817)	(0.000)	(0.270)	(0.986)
Points Exam 2	0.439***	0.428	0.296	0.361*	0.464***
	(0.000)	(0.103)	(0.101)	(0.054)	(0.000)
Female	−0.043	−0.034	−0.013	−0.020	−0.074**
	(0.277)	(0.210)	(0.685)	(0.605)	(0.045)
Controls	Yes	Yes	Yes	Yes	Yes
Class FE	Yes	Yes	Yes	Yes	Yes
Observations	160	77	83	90	70
Adjusted R^2	0.359	0.248	0.453	0.138	0.276

Note: This table . Dependent variable: percentage points exam 3. Covariates: percentage points exam 1, percentage points exam 2, gender, . The number of clusters is . p-values in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table 15: Class Treatment: Early math

Dep. Var: Test Scores	(1) All	(2) If Improved	(3) If Worsened	(4) If Better Half	(5) If Worse Half
Change Feedback	0.052	0.004	0.087^{***}	0.004	0.065
	(0.158)	(0.946)	(0.000)	(0.932)	(0.304)
Change × High Math Confi.	−0.025	0.006	−0.006	0.026	0.008
	(0.689)	(0.955)	(0.911)	(0.789)	(0.918)
Level Feedback	0.062	0.064	0.044^{**}	0.068	0.048
	(0.104)	(0.533)	(0.030)	(0.325)	(0.441)
Level × High Math Confi.	−0.038	−0.048	0.023	−0.050	0.009
	(0.460)	(0.617)	(0.836)	(0.674)	(0.892)
High Math Conf.	0.009	−0.027	0.003	0.014	−0.035
	(0.767)	(0.694)	(0.917)	(0.881)	(0.516)
Points Exam 1	0.362^{***}	0.331^{**}	0.467^{***}	0.368^{***}	0.399^{***}
	(0.000)	(0.011)	(0.000)	(0.000)	(0.000)
Points Exam 2	0.310^{***}	0.398^{***}	0.166	0.372	0.052
	(0.000)	(0.000)	(0.274)	(0.371)	(0.575)
Female	0.005	−0.007	0.020	−0.019	0.032
	(0.858)	(0.911)	(0.385)	(0.691)	(0.325)
Controls	Yes	Yes	Yes	Yes	Yes
Class FE	Yes	Yes	Yes	Yes	Yes
Observations	162	88	74	80	82
Adjusted R^2	0.515	0.418	0.599	0.277	0.555

Note: This table . Dependent variable: percentage points exam 3. Covariates: percentage points exam 1, percentage points exam 2, gender, . The number of clusters is . p-values in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table 15 shows that the response to feedback in the early treatment does not significantly depend on whether a student reported high or low math confidence. The negative sign and the magnitude of the coefficients of the interaction terms of change and level feedback with an indicator of high math confidence is a weak indication that especially students with low math confidence benefit from feedback, although none of the coefficients are significant

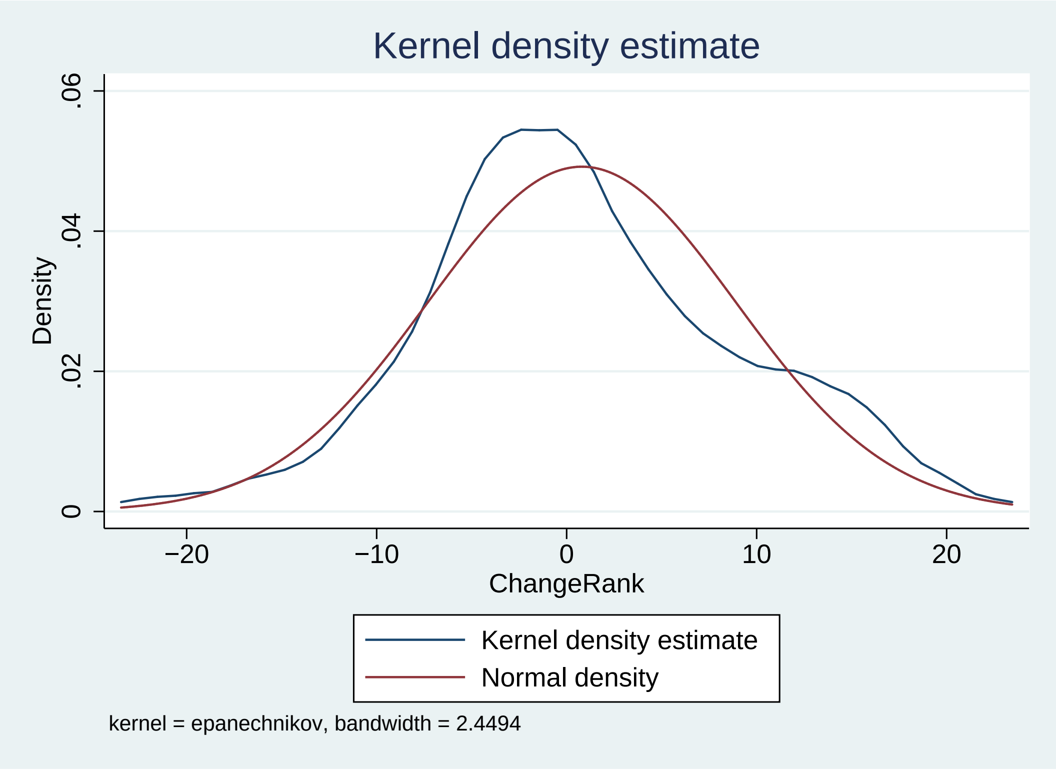
Table 16: Class Treatment: early selfesteem math

Dep. Var: Test Scores	(1) All	(2) If Improved	(3) If Worsened	(4) If Better Half	(5) If Worse Half
Change Feedback	0.109**	0.095	0.116**	−0.004	0.113**
	(0.042)	(0.483)	(0.027)	(0.931)	(0.049)
Change × High Self-Est.	−0.113**	−0.133	−0.066	0.006	−0.103
	(0.043)	(0.218)	(0.409)	(0.949)	(0.241)
Level Feedback	0.075**	0.072	0.061	−0.074	0.096**
	(0.025)	(0.530)	(0.189)	(0.153)	(0.010)
Level × High Self-Est.	−0.055	−0.066	−0.011	0.150**	−0.107**
	(0.192)	(0.537)	(0.859)	(0.026)	(0.031)
High Self-Est.	0.063**	0.051	0.065	−0.081	0.114***
	(0.018)	(0.673)	(0.407)	(0.107)	(0.000)
Points Exam 1	0.358***	0.353**	0.406**	0.412***	0.366***
	(0.000)	(0.053)	(0.018)	(0.000)	(0.000)
Points Exam 2	0.284**	0.331**	0.166	0.261	0.058
	(0.013)	(0.017)	(0.142)	(0.410)	(0.655)
Female	0.013	0.003	0.024	−0.035	0.044
	(0.648)	(1.000)	(0.384)	(0.369)	(0.179)
Controls	Yes	Yes	Yes	Yes	Yes
Class FE	Yes	Yes	Yes	Yes	Yes
Observations	162	88	74	80	82
Adjusted R^2	0.528	0.425	0.617	0.308	0.594

Note: This table . Dependent variable: percentage points exam 3. Covariates: percentage points exam 1, percentage points exam 2, gender, . The number of clusters is . p-values in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

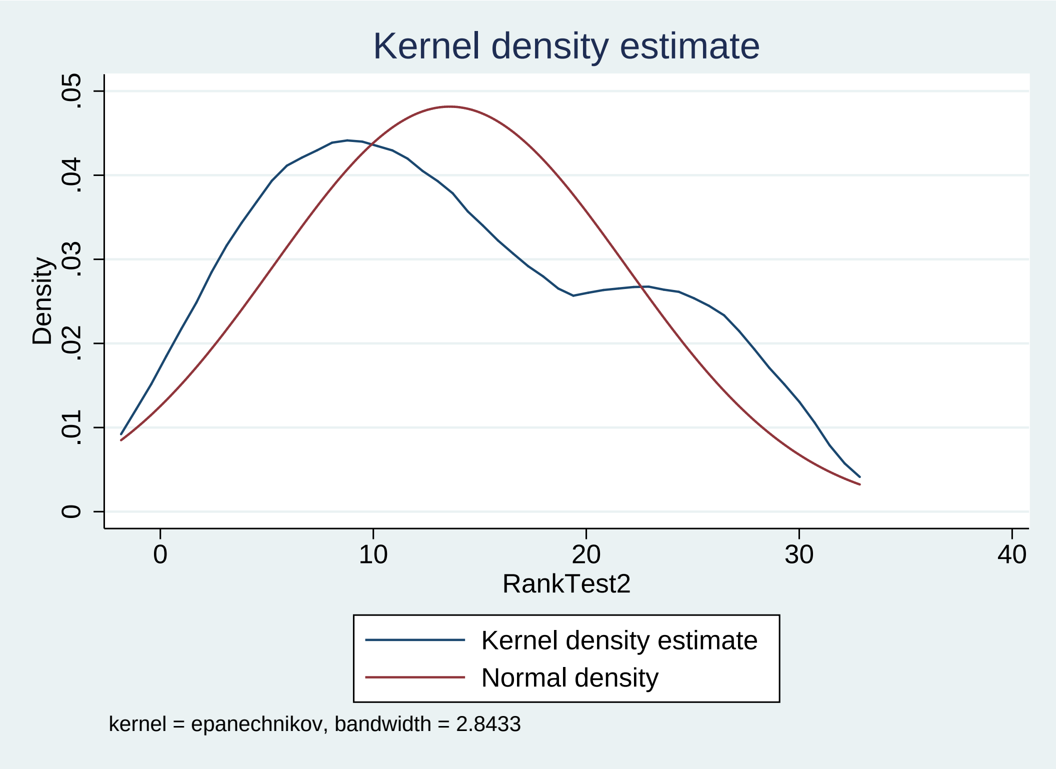
B Graphs

Figure 3: Feedback in Change Treatment



Note: This graph presents kernel density estimates for the feedback students received in the Change Treatment.

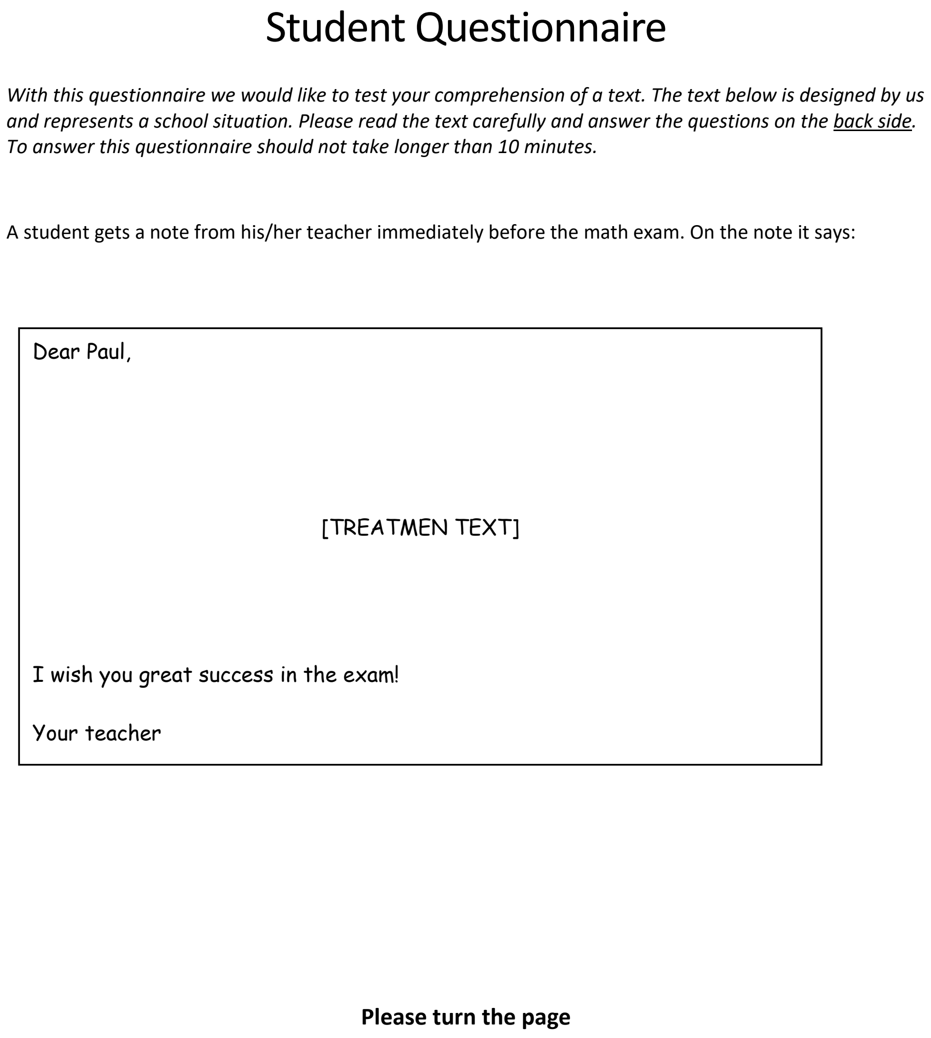
Figure 4: Feedback in Level Treatment



Note: This graph presents kernel density estimates for the feedback students received in the Level Treatment.

C Pretest

Figure 5: Pretest front and back [translated from German]



Student Questionnaire

With this questionnaire we would like to test your comprehension of a text. The text below is designed by us and represents a school situation. Please read the text carefully and answer the questions on the back side. To answer this questionnaire should not take longer than 10 minutes.

A student gets a note from his/her teacher immediately before the math exam. On the note it says:

Dear Paul,

[TREATMEN TEXT]

I wish you great success in the exam!

Your teacher

Please turn the page

1. Please summarize shortly (bullet points) what you have read on Pauls' note.

2. How do you think does Paul feel after reading the note?

1 □	2 □	3 □	4 □	5 □
very bad		medium		very good

3. How much do you think is Paul motivated to exert effort in the upcoming math exam?

1 □	2 □	3 □	4 □	5 □
not at all		medium		very strong

4. Paul was on rank 10 in the last exam. What is his rank now?

   a. There are 30 students in Pauls' class. How many children have a better rank than Paul in the second exam? [*Only 4a was asked in level feedback*]

5.

   a. [*change feedback:*]There are 30 students in Pauls' class and he ranked 10th in the last math exam. How did is rank change? (Please draw an error below)

   b. [*level feedback:*]There are 30 students in Pauls' class. How did Paul perform relative to the others? (Please mark the position with an X below)

the worst *the best*

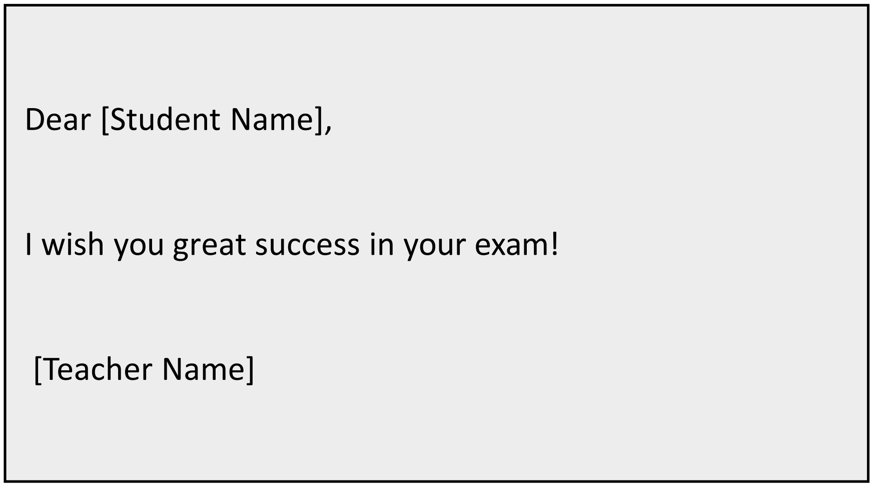
6. Do you know how many students are in your class?

   Number of students in your class:

Thank you very much

D Feedback Notes, Instructions, and Questionnaires

Figure 6: Feedback Note - Control Group [translated from German]

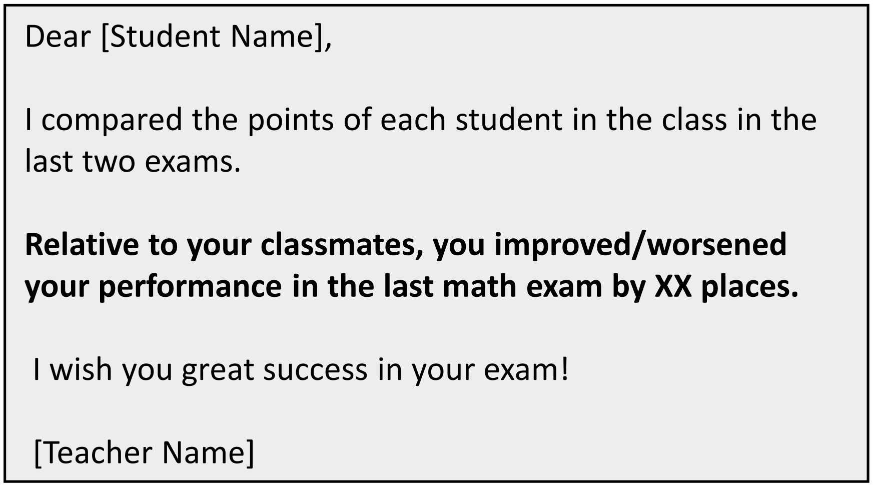


Dear [Student Name],

I wish you great success in your exam!

[Teacher Name]

Figure 7: Feedback Note - Change Treatment [translated from German]



Dear [Student Name],

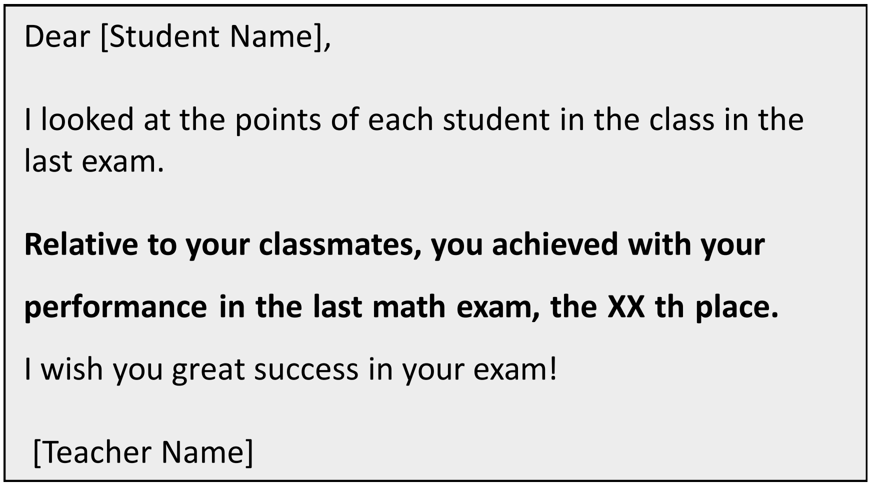
I compared the points of each student in the class in the last two exams.

Relative to your classmates, you improved/worsened your performance in the last math exam by XX places.

I wish you great success in your exam!

[Teacher Name]

Figure 8: Feedback Note - Level Treatment [translated from German]



Dear [Student Name],

I looked at the points of each student in the class in the last exam.

Relative to your classmates, you achieved with your performance in the last math exam, the XX th place.

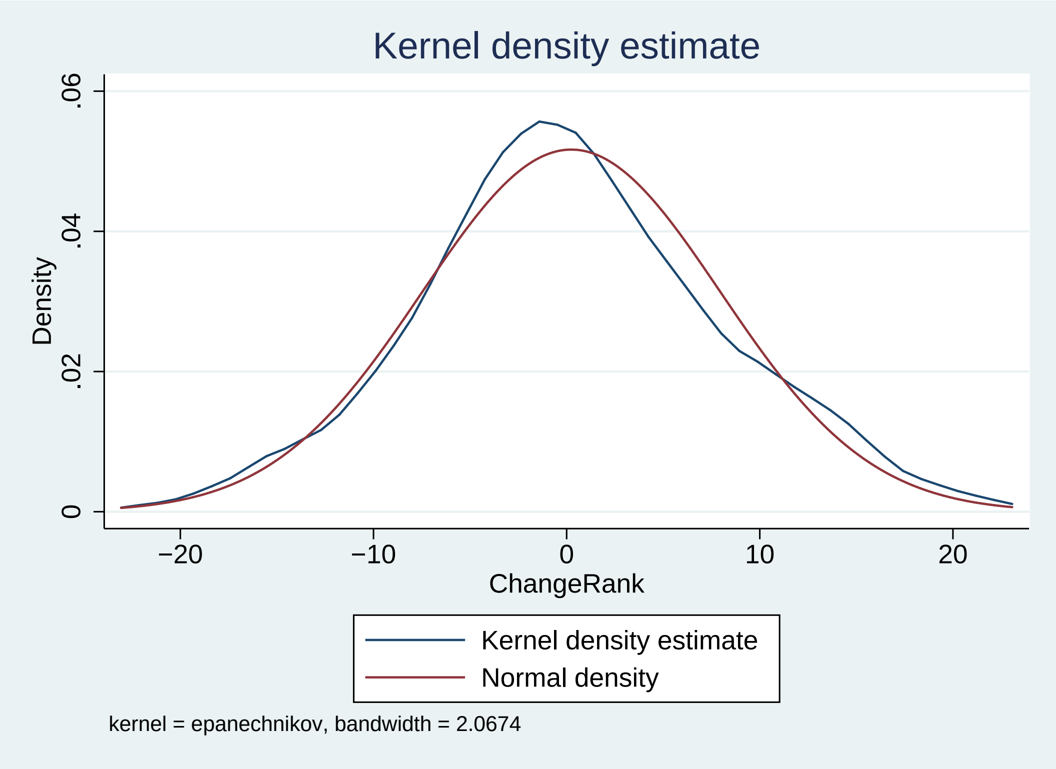
I wish you great success in your exam!

[Teacher Name]

Appendix - Not intended for Publication

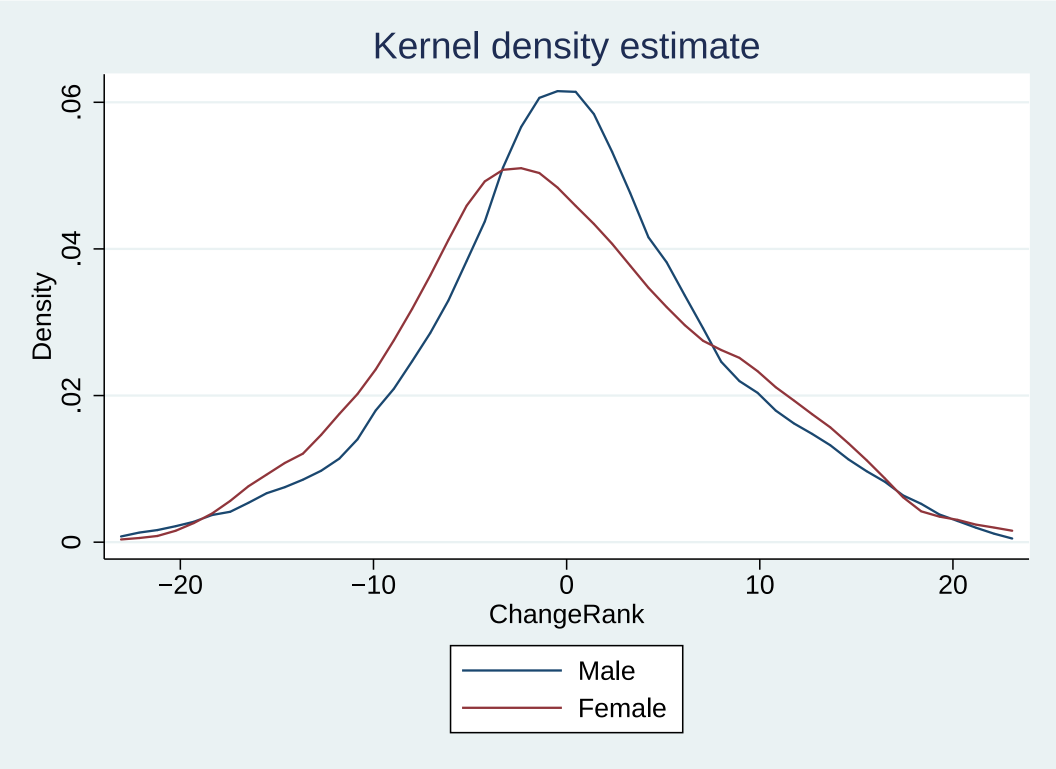
Kernel-density plots

Figure 9: Change in Rank



Note: This graph presents kernel density estimates for the change in rank between the first and the second exam.

Figure 10: Change in Rank by Gender



Note: This graph presents kernel density estimates for the change in rank between the first and the second exam separately for males and females.