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

Mira Fischer[†]and Valentin Wagner[‡]

1st August 2017

[PRELIMINARY DRAFT - PLEASE DO NOT CITE]

Abstract

We analyze the effects of relative performance information on high-stakes educational outcomes when the timing and the social reference frame of the feedback are manipulated. In a field experiment in secondary schools, students aged 10-11 years were provided with information about their absolute rank in the last exam, their change in ranks between the second last and the last exam, or with no feedback. With respect to timing, students received their feedback 1-3 days or immediately before the last exam of the semester. Overall, information about rank levels and rank changes both increased students' math performance when it was given 1-3 days before the exam and is most effective for students who recently suffered a decrease in their performance. In contrast, we find no significant effect of feedback given immediately before the exam. Moreover, we find that boys respond more strongly to early feedback than girls. Shedding light on potential mechanisms, we find that competitiveness does not explain our results while students with low self-esteem respond more strongly. Weak evidence suggests that early change feedback had a positive effect on students' belief in control over their outcomes.

Keywords: Timing of feedback, level feedback, change feedback, high-stakes test, field experiment, ranks

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, Henning Müller the participants of the 2nd Winter School on Applied Microeconomics, the 3rd Workshop on Educations Economics, the Workshop on Education, Skills, and Labor Market Outcomes in Oslo, as well as seminar participants at the University of Cologne, the University of Düsseldorf and the Norwegian School of Economics for helpful comments and suggestions. Particular thanks to all the teachers, parents and students who participated in the experiment. This research has been conducted with approval of the ethics committee of the University of Düsseldorf. The usual disclaimer applies.

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

[‡] University of Düsseldorf, 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 peer group is widely used and highly relevant in all contexts where the ability to motivate individuals is crucial, such as the work place, education, or sports. Feedback may be used in addition to incentive pay or when monetary rewards are too expensive, difficult to implement or socially unacceptable. Moreover, there are strong concerns that material incentives may crowd out intrinsic motivation (see e.g Deci et al. 2001). Information 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 studies find that it "works" (Azmat and Iriberri, 2010; Blanes i Vidal and Nossol, 2011; Tran and Zeckhauser, 2012; Azmat and Iriberri, 2016). 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] Therefore, determining which design features influence the success of feedback is an important object of study, which has not yet received as much attention as it deserves.

A broad categorization of feedback differentiates between directional feedback—providing clear directions on behaviors that the recipient should perform—and motivational feedback—providing information about rewards associated with particular behaviors (Ilgen et al., 1979). Economists have mainly studied feedback that falls into the latter category, either by providing a quantitative measure of past performance (outcome feedback, e.g. Tran and Zeckhauser, 2012; Azmat et al., 2016), such as a test score or a rank, or by allowing subjects to observe the behaviors of other people performing the same task (process feedback, e.g. Falk and Ichino, 2006; Mas and Moretti, 2009). Circumstantial variables are important for whether feedback is successful or not. For example, 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 giving rank information in public or private is more effective (Tran and Zeckhauser, 2012; Ashraf et al., 2014; Gill et al., 2016; Hannan et al., 2013; Tafkov, 2013). The effectiveness of feedback also depends on whether a pay-for-performance or a flat incentive scheme is present (Azmat and Iriberri, 2016), or whether the feedback is sufficiently precise (Hannan et al., 2008).[2] Overall, the question of what makes feedback effective has received rather little attention from economists. Evidence on other potentially important features of feedback—e.g. how it is timed and which reference frame it uses—is scarce, although they could be crucial for its success.

This paper begins to fill this gap by presenting results of a field experiment in secondary schools in which we exogenously vary *whether* students receive private rank feedback, *when* they receive it and *how* it compares students (reference frame). To study the effects of timing and reference frame of relative feedback on high-stakes educational outcomes, 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 of three mathematics exams of the school semester, which was thus crucial for progression to the next grade. With regard to the reference frame, students within the same class either received feedback about (i) their performance rank in the last exam, (ii) their change in rank between the second last and the last exam, or (iii) no feedback.

We find that feedback is only effective to increase subsequent performance when given a few days before the last exam and that both change and level feedback work equally well to increase performance. In early treatment classes, students receiving feedback about their rank level significantly increase their performance

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

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

by 3.9 percentage points compared to students receiving no feedback, while students receiving feedback about rank changes significantly increase their performance by 3.8 percentage points.[3] These effects are driven by giving feedback to those students who recently suffered a decrease in their performance. Furthermore, boys respond more strongly to feedback than girls. In contrast, any feedback given to students immediately before the exam tends to lower subsequent performance by 0.2 percentage points with change feedback and 2.2 percentage points with level feedback but the estimated impacts on exam scores are not significant.

To understand the mechanisms that drive the effects of feedback on performance, we elicit students' belief in the effectiveness of their effort and use validated scales to measure their character traits, such as competitiveness, confidence in academic abilities, locus of control, and self-esteem.We find that competitiveness, confidence, and locus of control do not interact with our feedback intervention but that 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. Moreover, we find weak evidence that the change feedback given early had a positive effect on students' belief that they could affect their outcomes by exerting effort.

We made careful choices on our intervention's design features and target group. Feedback was given at different points in time because outcomes in the workplace or educational settings may be influenced by different types of effort exerted at different times – preparation effort and effort at the task itself.[4] While earlier feedback may 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 performance. Although there might be a tendency to try to reinforce effort by giving feedback after observing high performance or personal improvement, it could be more motivating to receive feedback after low performance or slacking off. 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.[5]

Students were provided with relative performance information as people are strongly motivated by it, 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).[6] However, rank feedback that compares 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.[7] This feature of rank feedback might play a role in education in particular because 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) and 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. Feedback about how one's performance has changed in the past may also help to promote the belief that skills can be developed by exerting effort

[3] The difference between change and level feedback is statistically not significant.

[4] See Levitt et al. (2016a); Wagner (2016) for studies disentangling the effort from the learning effect in educational settings.

[5] Positive feedback may build self-efficacy, i.e. an individual's belief that he or she has the skills necessary to complete a particular task (Bandura, 1986), and reinforces behavior. However it can also result in reduced effort or motivation (Cianci et al., 2010), while failure can be a strong motivator (Fishbach et al., 2010).

[6] In the context of education, ability relative to one's peer-group seems to be particularly important for motivation. Murphy and Weinhardt (2014) and Elsner and Isphording (2017) have found that, holding ability constant, a higher rank within one's class or cohort lead to more favorable beliefs in own ability and better academic outcomes later in life. Already Adam Smith thought that "rank among our equals, is, perhaps, the strongest of all our desires" (Smith, 1759).

[7] For example, it might result in receiving only negative outcome feedback despite individual progress, which could result in learned helplessness, a belief that one has little to no control over one's outcomes, no matter how hard one tries (Seligman, 1975).

(also called a "growth mindset" in the psychological literature, see Paunesku et al. 2015; O'Rourke et al. 2014). For these reasons, relative 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 academic skills.

We study students early in their academic career because 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). Because of 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 (Cunha and Heckman, 2007). In light of these facts the question of whether feedback in schools can be improved by deliberately selecting it$s$ references frame as well as its timing is an important one.

To our knowledge, this is the first study that varies the timing of feedback and the first to compare two generic types of relative feedback (about levels and changes of performance). Furthermore, it tests feedback on a sample of secondary school students (aged 10 - 11 years), while so far researchers have exploited data of natural experiments with older students or tested feedback on university students. Our results are not only relevant for educators but the general findings extend to other settings where feedback is given with the intention to increase motivation.

The remainder of this paper is organized as follows. The next section gives a brief overview of the related literature. In section 3 we report the results of a pretest. Section 4 describes our experimental procedures. Section 5 presents the results and potential mechanisms driving our results are discussed in section 6. Section 7 concludes.

2 Related Literature

Besides screening for talent[8], economists traditionally focus on the introduction of incentives 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.[9] 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.

This paper is related to the strand of literature analyzing the effectiveness of non-monetary incentives and behavioral interventions on performance in educational settings. Few studies so far have looked at the effects of feedback in the context of education and among those all but one (Azmat and Iriberri, 2010) have relied on university student samples. Tran and Zeckhauser (2012) provide Vietnamese students participating in an English-testing experiment either with private feedback (by phone) or private plus public feedback (postings on the university's noticeboard and website) about their ranking in in-course mock exams. Overall,

[8] Surprisingly little of the large heterogeneity of teacher effectiveness can be explained by observable teacher characteristics (Hanushek and Rivkin, 2006), which makes it hard to improve educational outcomes by screening for good teachers.

[9] Damgaard and Nielsen (2017) recently review the use of behaviorally motivated interventions in education.

the authors find a positive effect of feedback on the final English test and that private plus public feedback tends to outperform private feedback alone. This difference, however, was only marginally significant.[10] A more recent study by Bandiera et al. (2015) exploits 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. Azmat et al. (2016) provided college students with feedback on their position in the grade distribution 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 these studies 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 and there is—to our knowledge—no *randomized controlled* field experiment on the effectiveness of performance feedback on educational outcomes of children. Azmat and Iriberri (2010) study the motivational effect of relative performance feedback among high school students in Spain (aged 14 - 18) in a natural field experiment. For one school year, a high school in the Basque Country adopted a new system of producing report cards providing students with information on whether they were performing above or 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.

The paper by Azmat and Iriberri (2010) is the one most similar to ours with respect to the population studied, although their student sample is about 3 - 8 years older than ours. With respect to the dimensions of feedback – timing and social reference frame – manipulated in our design, we are not aware of any similar studies.

3 Pretest

Teachers in Germany often provide their students with a statistic about the frequency of grades in their class *after* an exam. Students therefore have some imprecise information about how their performance compares to the performance of other students. Students in our sample are quite young and in order to test if they understand our feedback (to disentangle lack of understanding and ineffectiveness of feedback) and how they interpret it (to enable us to interpret possible effects), we conducted a survey before implementing the field experiment in 4 classes of 6 schools with a total of 151 students of the same age group. This was a convenience sample gathered through personal contacts.

The survey consisted of a two-page questionnaire.[11] On the first page students saw a feedback note of a fictitious student named "Paul" and were asked to imagine themselves in his position. On the back of the page, students had to shortly summarize the information of the first page and answer a quiz to test whether they understood it correctly. They were also asked to give their guess of how Paul feels (very good - very bad) after having read the feedback note and of how much (not at all - very strongly) Paul will be motivated to exert effort in the next exam. We also asked students whether they knew the size of their class, which is crucial for correctly interpreting rank feedback. We randomly varied Paul's feedback and presented students

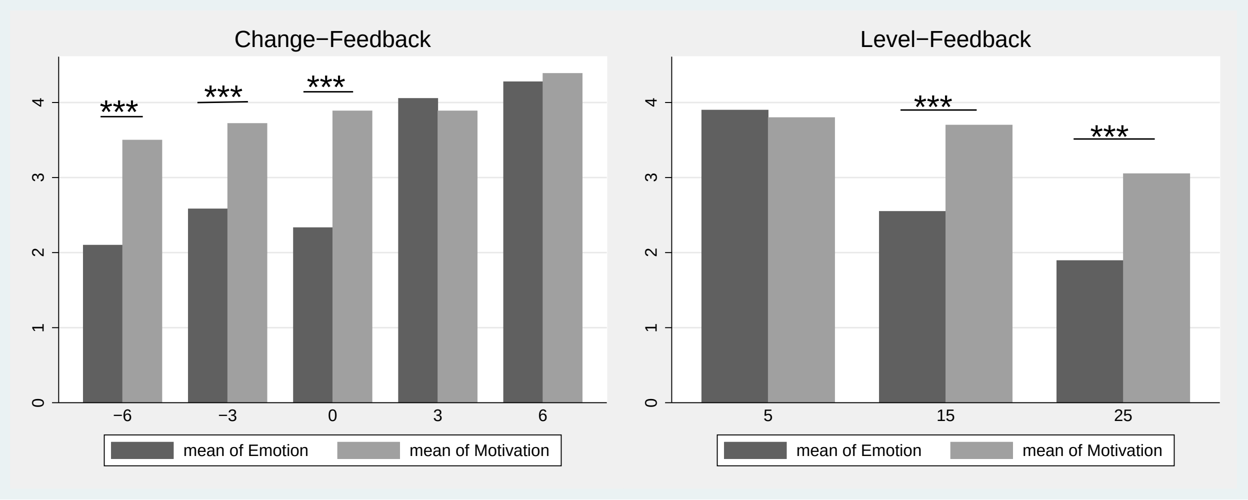
[10] In contrast to Tran and Zeckhauser (2012), Ashraf et al. (2014) find that private plus public feedback reduces performance of health workers in Zambia in a nationwide training program.

[11] See Appendix C.

either with change feedback or level feedback. Furthermore, we varied the size of the change in ranks (-6, -3, 0, 3, or 6) and the rank levels (5, 15, or 25).

Overall students seem to understand the feedback notes. 85.56% of the students could correctly calculate by how much Paul's rank changed and 94.74% could correctly determine the position of Paul's rank when given level feedback. Moreover, 86.09% of the students know the exact size of their class. The mean responses to the questions concerning Paul's emotions and motivation are presented in Figure 1 and differed by the reference frame of feedback. On average, students thought that Paul would be more motivated when receiving feedback about changes than when receiving level feedback (3.87 vs 3.53, p = 0.0745) while students do not think that the two references frames of feedback differently affect emotions (3.05 vs 2.80, p = 0.2353). Importantly, the significant difference in reported motivation between the change feedback and the level feedback may be driven by the chosen ranks and change in ranks and might become insignificant when testing different numbers. However, regardless of whether feedback is given in levels or in changes, Paul's emotions and motivation are rated more positively with higher rank levels and positive rank changes. Interestingly, students believe that negative feedback (negative change in ranks or rank level below median) will make Paul feel worse than positive feedback but his motivation to exert effort is rated approximately the same with negative and positive feedback.

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



Note: This graph shows the results of a pretest separately for change feedback (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 to 5 scale. Feedback notes in the pretest were varied such that students faced either a change in Paul's rank of -6, -3, 0, 3 or 6 or the ranks 5, 15 or 25. Differences between emotions and motivation were tested with a mean-comparison tests.

4 Experimental Intervention

The experiment was conducted in 7 secondary schools with in total 19 classes (grades 5 and 6) in the western German cities of Bonn, Cologne, and Düsseldorf and was approved by the ethics committee of the

University of Düsseldorf. 352 students of grades five and six received parents' consent (about 73,74% per class) participated during May and June 2016.[12] Researchers were never present in the classroom to maintain a natural examination situation and the feedback was given to students by their math teacher to maximize its credibility.[13] 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 procedure of the intervention 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 templates for providing results of the two prior math exams, including the classes' grades and points in each of the two exams and the maximum number of points reachable.[14] For those students whose parents consented, teachers provided us with names, which enabled us to print personalized feedback notes by calculating students' ranks in the last math exam and their change in ranks from the second last to the last math exam. The feedback notes were folded and students' names were written on the front.[15] The second envelope, containing the feedback notes, instructions, a result template for the third exam, as well as student questionnaires, was sent to schools timely within a few days before the third exam, which was also the last of three exams during the semester, and thus implied high stakes for the students.

Treatments We want to test how relative performance feedback affects a student's performance in a high-stakes math exam. As described above, relative feedback has often proven effective in raising performance (Azmat and Iriberri, 2010; Blanes i Vidal and Nossol, 2011; Tran and Zeckhauser, 2012; Azmat and Iriberri, 2016) but has also been found to backfire (Azmat et al., 2016; Bradler et al., 2016; Ashraf et al., 2014; Barankay, 2012) and there is little evidence on the effects of feedback in schools. Rank feedback also seems promising in light of recent findings that a student's rank within their class or cohort affects later achievement independently of underlying ability (Elsner and Isphording, 2017; Murphy and Weinhardt, 2014).[16] Based on a 2 X 3 design, we vary both the *timing* of feedback and the *reference frame* of feedback independently. We are not aware of any studies that have looked at the effect of timing, although it is potentially very important because test outcomes are influenceable both by learning and test taking effort, exerted at different times. Furthermore, feedback can be given in terms of individual levels of performance (rank in last test) and in terms of changes of performance (e.g. change in rank between second last and last math test). While all prior studies on rank feedback have used levels, the tournament literature points towards this being harmful in settings where ability differences are large (Gürtler and Harbring, 2010), such as in many classrooms. There is also evidence from the psychological literature that promoting the belief that own skills are changeable improves a student's motivation (Paunesku et al., 2015; O'Rourke et al., 2014).

The timing of feedback was randomized on class level. Students either received feedback 1-3 days before the exam (Early Timing) or immediately before the exam sheets were handed out (Late Timing). The reference frame of feedback was randomized on student level. Within the same class, students with parents'

[12] We contacted 142 secondary schools in the federal state of 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.

[13] The credibility of the source has a substantial effect on how feedback is interpreted. Ilgen et al. (1979) identified two components of source credibility: expertise and trustworthiness.

[14] See Appendix X for teacher instructions.

[15] 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 students 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.

[16] Murphy and Weinhardt (2014) find that students with a one standard deviation higher rank in primary school will score 0.08 standard deviations better at age 14 and Elsner and Isphording (2017) find that high school students with a higher rank have higher expectations about their future career outcomes, are more optimistic and self-confident and increases the likelihood of going to college.

permission to participate received personalized written feedback about their rank in the last math exam (LEVEL FRAME), about their change in rank between the second last and the last math exam (CHANGE FRAME), or a personalized note that only wished them good luck (CONTROL). In all treatments, teachers gave a folded feedback note to each student that had the student's name written on its outside. To personalize the feedback, the note addressed the student by their first name and was signed with the teacher's name (see Appendix D for English translation of the exact wording and layout of the notes). While students in CONTROL received no information about their past performance, in CHANGE FRAME, students received information about their change in rank between the two previous exams but no information on their absolute ranks in these tests ("*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*"). Students in LEVEL FRAME were informed about their relative rank in the last exam but received no information on their performance in the second last exam or about how their performance changed ("*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*"). As students had received their grades in the last two exams after the teachers had graded them (i.e. approximately 2 and 4 months before the last exam, respectively), the feedback information served as a reminder that contained more detailed information about different aspects of their relative performance.

In EARLY TIMING, students had to fill in a questionnaire immediately after receiving the feedback notes, while in LATE TIMING students had to fill in a questionnaire immediately after completing the exam. Due to time constraints, in LATE TIMING, the questionnaire was shorter and did not include all scales included in the EARLY TIMING questionnaire. The questionnaire elicited effort-effectiveness beliefs, non-cognitive skills, character traits, and demographic information. It enables us to study whether the feedback possibly affects test outcomes by changing beliefs about how easily outcomes can be affected by effort. Furthermore, feedback can possibly have heterogeneous effects on students with different gender (Buser and Yuan, 2016), non-cognitive skills and character traits . The questionnaire gives us the possibility to explore these possible differences, although some of the scales might be crude measures of the underlying traits. Questions on non-cognitive skills 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 (based on Bandura 1986, 1997) and self-esteem (German version of the Rosenberg self-esteem scale, Collani, Herzberg 2003).

After students filled in the questionnaires, teachers collected them, while students were required to crumble the feedback notes and throw them in a garbage bin. Upon sending the results of the final exam as well as the filled-in questionnaires, teachers were asked to fill in a short online survey.

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

Blocked on school level, classes were randomized either into the LATE TIMING treatment or the EARLY TIMING treatment. With respect to these class-level treatments (EARLY TIMING, LATE TIMING) non-random self-selection was possible as parents learned whether feedback will be given 1-3 days before the exam or immediately before the exam in their child‘s class. This was necessary to receive parents' fully informed consent. Within classes, students were then randomized into the Control group, Change Frame treatment or Level Frame treatment. Parents did not learn into which of the three treatments their child was allocated as randomization into student-level treatments took place only after we obtained parents' consent. Hence, non-random self-selection into the student-level treatments was not possible.

Overall, randomization was successful as no significant differences between treatments are found in any important dimensions (prior test scores and grades, gender, student demographics). In the following we will discuss the randomization checks in detail. Table 6 in Appendix A reports differences between Early Timing and Late Timing. Student and teacher observables do not differ significantly between these class-level treatments, except with respect to the share of students per class who participated and teacher experience. Fewer students per class participate in Early Timing as compared to Late Timing and teachers in Early Timing are more experienced than teachers in Late Timing.

Surprisingly, the share of participants turned out to be significantly lower in the Early Timing treatment compared to the Late Timing treatment. We expected the opposite as parents might be concerned about larger negative (emotional) effects on exam outcomes of their children when feedback is 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 treatment and both Change Frame and Level Frame treatment in measures of past performance (ranks in exam 1 and 2, points in exam 1 and 2, change in rank and share of worseners).[17] We can also check whether students 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 Timing treatment: 2.61 vs 2.70; Early Timing treatment: 2.90 vs 3.07) and exam 2 (Late Timing treatment: 2.59 vs 2.86; Early Timing treatment: 2.60 vs 2.82).

As the treatment groups are balanced on student characteristics, we do not expect teacher experience to influence our results. Teacher characteristics, such as education or experience, do not explain much of the variation in educational outcomes (Rivkin et al., 2005). Moreover, our analysis controls for teacher grading by accounting for prior test scores and by standardizing test scores on class level.

Tables 7 - 9 in Appendix A present randomization checks for student-level treatments (Change Frame, Level Frame, Control), pooled and separately for each class-level treatment. As mentioned above, self-selection into these treatments was not possible, as students had no information on assignment prior to the intervention, and student observables in the student-level treatments are not significantly different from each other in any important dimension. However, we find small statistical differences in two survey questions. Students' beliefs about their parents' valuation of education differs slightly between the Level Frame and the Change Frame treatments while students' reported books at home differs slightly between the Change Frame and the Control treatments Nevertheless, these differences are small and should not affect our results, all the more so as prior test scores (our main outcome variable) do not differ significantly between any treatment groups.

To summarize, students do not differ in any important dimensions across student-level treatments. On

[17] Overall, 157 (30.84%) students did not get their parents' consent to participate in the experiment [68 (28.10%) in the Late Timing treatment and 89 (33.33%) in the Early Timing treatment]. In 16 out of 19 classes, more than 50% of the students within the class participated.Participation rates ranged from 37.93% - 100% on class level. We can check whether students 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 in both exam 1 (Late Timing treatment: 2.61 vs 2.70; Early Timing treatment: 2.90 vs 3.07) and exam 2 (Late Timing treatment: 2.59 vs 2.86; Early Timing treatment: 2.60 vs 2.82)

class-level treatments, fewer students participate in the Early Timing 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 consist of pre and post intervention performance measures provided by the teachers as well as information from student questionnaires. Importantly, we have very detailed information on students' past performance as we know students' grades and points in the two last exams before the interventions as well as the maximum score possible in the exams. This data can be treated as exogenous in the analysis because the information was given to students several months before teachers learned about the study 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 students 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 1 to 6, where 1 is the highest and 6 is the lowest grade.[18] Table 1 summarizes the feedback students received by treatment and reveals that students, in part, received a strongly negative or strongly 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 Frame	Early Timing	59	0.763	8.052	-21	+21
	Late Timing	57	0.842	8.239	-19	+19
Level Frame	Early Timing	64	13.922	8.407	1	30
	Late Timing	60	13.233	8.208	1	30

Note: This table presents descriptive statistics of the feedback given to students by class-level and student-level treatments.

5.3 Impact of feedback on performance

In the following we present our results. We first analyze the effect of *timing* of feedback (Early Timing versus Late Timing) on performance, which was randomized at the class level. Then we will analyze the overall effect of the *reference frame* of feedback (Change Frame versus Level Frame versus Control), 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 valence of feedback. The following tables present results from linear regressions (OLS) that include prior performance as linear control variables and student characteristics as control variables, as well as a constant. Furthermore, regressions analyzing treatments that were randomized onstudent level contain class fixed effects and analyzing treatments randomized on class level include class characteristics as control variables and contain school fixed effects.The advantage of including class fixed effects is... In each case the reported standard errors are corrected using Biased-Reduced Linearization

[18] 1.0 to 1.3 =A;>1.3 to 2.3=B; >2.3 to 3.3=C; >3.3 to 4.0= D; >4.0=F

(BRL) (Bell and McCaffrey, 2002), which is robust to both heteroscedasticity and small sample size and clustered at the class level. First, we study the effect of *timing* of feedback on performance to learn whether students receiving the intervention 1-3 days before the exam had different outcomes than students receiving the intervention immediately before the exam. Then, we will look that the Early Timing and the Late Timing group separately to study the effect of *reference frame* of feedback. This will allow us to explain whether a possible difference between the Early Timing and the Late Timing groups is driven by the (differential) effects of the Change Frame, or the Level Frame feedback, or by both.

5.3.1 The role of timing

To analyze whether students in the Early Timing or Late Timing groups had different outcomes in the final exam, we estimate the following linear model (OLS)[19]:

$$\begin{aligned} PointsTest3_i = \beta_0 + \beta_1\, EarlyTiming_i + \beta_2\, PointsTest1_i + \beta_3\, PointsTest2_i + \\ \gamma\, ClassCovariates_i + \mu\, StudentCovariates_i + \delta\, School_i + \varepsilon_i \end{aligned} \tag{1}$$

$PointsTest3_i$ are the percentage points in the final math exam of student i, $PointsTest1_i$ and $PointsTest2_i$ are the percentage points in the second last and the last exam of student i, $StudentCovariates_i$ is a vector of characteristics of student i: student i's gender, whether student i has a foreign sounding name (to capture migration background), whether student i has siblings, and whether student i has an own room.$ClassCovariates_i$ is a vector of characteristics of the student's class: the teacher's gender and the class size. $School_i$ controls for school fixed effects and ε_i is a stochastic i.i.d. error term.

Table 2 reports on the effect of receiving feedback 1-3 day prior to the exam pooled by student-level treatment. Column 1 presents the results for the whole sample, while columns 2 and 3 split the sample by improvement/worsening from second last to last test and columns 4 and 5 split the sample by better/worse half in terms of ranks in the last test. 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 feedback to students who decreased their relative performance prior to the exam we study.[20] However, as can be seen in models 4 and 5, both students who had an above and those who had a below median rank seem to react to the feedback approximately similarly. Ability level does not seem to explain why students who received the feedback earlier had better test outcomes than students who received the feedback later.

[19] Regressions without control variables can be found in Table 10 in the Appendix.

[20] This result cannot be driven by regression to the mean as the model compares students in both treatments who got worse among each other.

Table 2: Early Timing vs. Late Timing

Dep. Var: Test Scores	(1) All	(2) If Improved	(3) If Worsened	(4) If Better Half	(5) If Worse Half
Early Timing	0.056*	0.026	0.089***	0.059	0.065*
	(0.031)	(0.044)	(0.029)	(0.045)	(0.033)
Points Exam 1	0.254***	0.177	0.519***	0.263***	0.238***
	(0.073)	(0.170)	(0.100)	(0.092)	(0.074)
Points Exam 2	0.352***	0.454***	0.125	0.315**	0.253***
	(0.081)	(0.158)	(0.097)	(0.137)	(0.081)
Female	−0.020	−0.019	−0.013	−0.023	−0.013
	(0.019)	(0.031)	(0.022)	(0.021)	(0.027)
SchoolFE	Yes	Yes	Yes	Yes	Yes
Pupil Control	Yes	Yes	Yes	Yes	Yes
Class Control	Yes	Yes	Yes	Yes	Yes
N	319	163	156	168	151
adj. R^2	0.390	0.291	0.507	0.207	0.320

Note: This table presents the effect of feedback timing on performance in the last exam using a linear regression model including school fixed effects. Column 1 presents the results for the whole sample, columns 2 and 3 split the sample by students that could improve/worsen their rank from second last to last exam and columns 4 and 5 split the sample by students in the upper/lower half in terms of ranks in the last exam. Dependent variable: percentage points exam 3. Covariates: gender teacher, class size, percentage points exam 1, percentage points exam 2, gender, own room, foreign name, siblings. Standard errors are reported in parentheses, clustered on classroom level and corrected using biased-reduced linearization. The number of clusters is 19. * 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[21] or (iv) a negative emotional effect of students in the Late-Treatment.[22] 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 11 in the Appendix, 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

[21] 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.

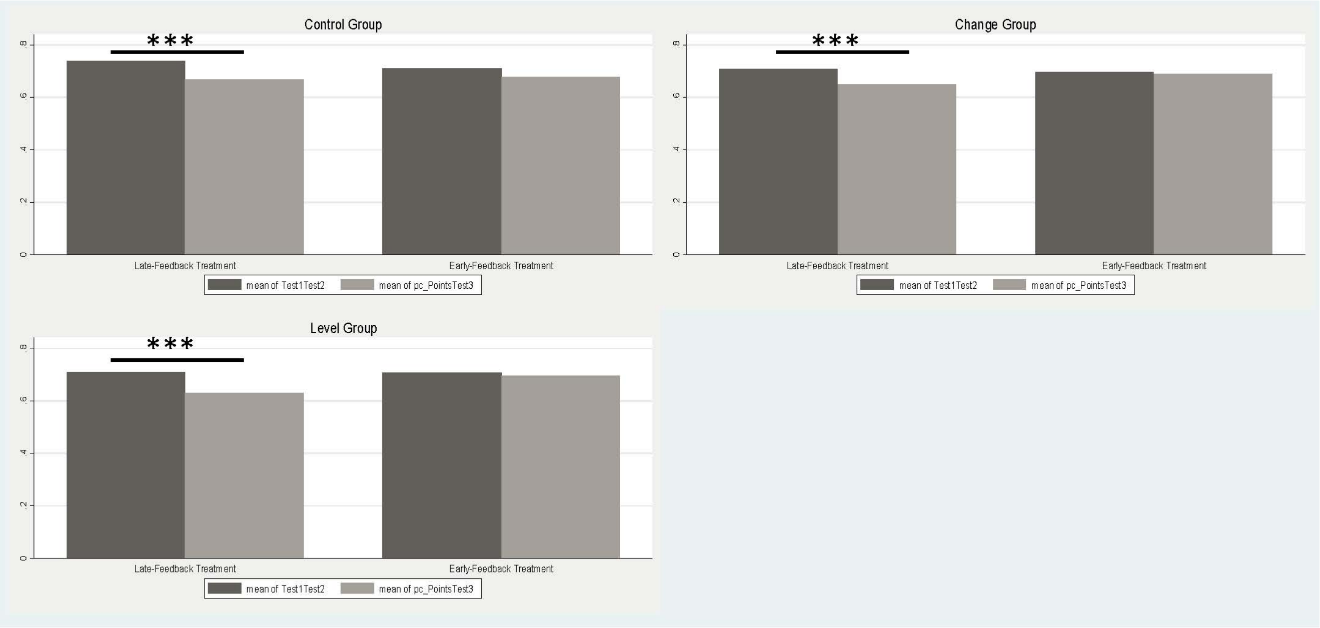
[22] We do not claim that (i)-(iv) are the only candidates which could explain the difference between the two class-level treatments but think that they are the most likely.

effect of answering the questionnaire.

Differences between the Late Timing and Early Timing treatments could be due to an increase in performance of students in Early Timing or due to a decrease in performance of students in Late Timing. 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" pattern. 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 who receive feedback 1-3 days before the exam perform better than students who receive feedback immediately before the exam.*

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 student-level treatment.

5.3.2 The role of reference frame of feedback

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

$$PointsTest3_i = \beta_0 + \beta_1\, ChangeFeedback_i + \beta_2\, LevelFeedback_i + \beta_3\, PointsTest1_i + \beta_4\, PointsTest2_i$$

$$+ \gamma\, Covariates_i + \delta\, Class_i + \varepsilon_i \qquad (2)$$

$PointsTest3_i$ are the percentile points in the final math exam of student i, $PointsTest1_i$ and $PointsTest2_i$ are the percentile points in the second last and the last exam of student i, $Covariates_i$ is a vector of characteristics of student i: students i's gender, whether student i has a foreign sounding name (to capture migration background), whether student i has siblings, and whether student i has an own room.$Class_i$ controls for class fixed effects and ε_i is a stochastic i.i.d. error term.

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.

Change and level feedback given early Table 3 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 3: Change Frame vs. Level Frame vs. Control - Class-Level Treatment: Early Timing

	(1)	(2)	(3)	(4)	(5)
Dep. Var: Test Scores	All	If Improved	If Worsened	If Better Half	If Worse Half
Change Frame	0.038^{*}	0.002	0.081^{***}	0.021	0.074^{**}
	(0.022)	(0.051)	(0.027)	(0.030)	(0.037)
Level Frame	0.039^{**}	0.026	0.053^{**}	0.034	0.056^{*}
	(0.016)	(0.037)	(0.025)	(0.026)	(0.034)
Points Exam 1	0.358^{***}	0.318^{**}	0.473^{***}	0.404^{***}	0.385^{***}
	(0.046)	(0.149)	(0.127)	(0.082)	(0.054)
Points Exam 2	0.297^{***}	0.350^{***}	0.161	0.324	0.037
	(0.067)	(0.128)	(0.121)	(0.324)	(0.085)
Female	0.005	−0.006	0.020	−0.025	0.032
	(0.029)	(0.051)	(0.022)	(0.037)	(0.030)
ClassFE	Yes	Yes	Yes	Yes	Yes
Pupil Control	Yes	Yes	Yes	Yes	Yes
N	160	87	73	79	81
adj. R^2	0.517	0.426	0.612	0.301	0.556

Note: This table presents the effect of change frame and level frame feedback when given 1-3 days in advance using a linear regression model including class fixed effects. Column 1 presents the results for the whole sample, columns 2 and 3 split the sample by students that could improve/worsen their rank from second last to last exam and columns 4 and 5 split the sample by students in the upper/lower half in terms of ranks in the last exam. Dependent variable: percentage points exam 3. Covariates: percentage points exam 1, percentage points exam 2, gender, own room, foreign name, siblings. Standard errors are reported in parentheses, clustered on classroom level and corrected using biased-reduced linearization. The number of clusters is 19. * p<0.10, ** p<0.05, *** p<0.01.

Table 4: Change Frame vs. Level Frame vs. Control - Class-Level Treatment: Late Timing

	(1)	(2)	(3)	(4)	(5)
Dep. Var: Test Scores	All	If Improved	If Worsened	If Better Half	If Worse Half
Change Frame	−0.002	0.022	−0.029	0.010	0.011
	(0.018)	(0.040)	(0.023)	(0.033)	(0.038)
Level Frame	−0.022	−0.009	−0.023	−0.036	−0.020
	(0.020)	(0.031)	(0.046)	(0.026)	(0.038)
Points Exam 1	0.125	0.105	0.382^{***}	0.183	0.008
	(0.122)	(0.293)	(0.129)	(0.131)	(0.182)
Points Exam 2	0.437^{***}	0.429	0.256^{*}	0.339^{***}	0.475^{***}
	(0.110)	(0.269)	(0.137)	(0.095)	(0.121)
Female	−0.041	−0.047	−0.021	−0.022	-0.065^{*}
	(0.031)	(0.039)	(0.028)	(0.035)	(0.035)
ClassFE	Yes	Yes	Yes	Yes	Yes
Pupil Control	Yes	Yes	Yes	Yes	Yes
N	159	76	83	89	70
adj. R^2	0.361	0.205	0.456	0.155	0.307

Note: This table presents the effect of change frame and level frame feedback when given immediately before the exam using a linear regression model including class fixed effects. Column 1 presents the results for the whole sample, columns 2 and 3 split the sample by students that could improve/worsen their rank from second last to last exam and columns 4 and 5 split the sample by students in the upper/lower half in terms of ranks in the last exam. Dependent variable: percentage points exam 3. Covariates: percentage points exam 1, percentage points exam 2, gender, own room, foreign name, siblings. Standard errors are reported in parentheses, clustered on classroom level and corrected using biased-reduced linearization. The number of clusters is 19. * p<0.10, ** p<0.05, *** p<0.01.

Change and level feedback given late Table 4 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, self-esteem matter for how a child reacts to feedback? 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 gender and certain subgroups. 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.

Gender 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 12 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.

Character traits The psychological literature suggests that individual differences matter for how people react to (positive and negative) feedback (Ilgen et al., 1979). For example people with a more external locus of control may think that a bad outcome is due to factors they cannot control and may therefore not react to negative feedback by increasing their effort (Lam & Schaubroeck, 2000). People with low self-esteem might have a more negative emotional response to negative feedback than people with higher self-esteem and may therefore be more distracted from a subsequent task (Fedor, Davis, Maslyn, & Mathieson, 2001). People with high self-efficacy, i.e. a strong belief that they have the skills to complete a particular task, have been found to be more motivated by feedback than people who have low self-efficacy (Colquitt, LePine, & Noe, 2000; London, 2003). 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. 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.

Beliefs about effectiveness of learning effort Turning to beliefs about the effectiveness of effort, we find some weak evidence that the change feedback in the early treatment has a positive effect on the belief that one's outcomes can be affected by one's effort. Students who received change feedback in the early treatment have a 0.17 standard deviations higher belief that they will achieve a higher grade if they exert more effort at studying than the control group. This effect is significant at the 10 percent level. Interestingly, this effect seems to be homogeneous for both positive and negative change feedback, although we have to caution that the coefficients in the subgroup analyses do not turn out significant. At the same time, the coefficient of the level feedback is very close to zero and highly insignificant. In the late treatment we do not find feedback to affect the belief in the effectiveness of effort overall. Interestingly, however, while the coefficient of the level feedback is again close to zero, the coefficient of the change feedback, although insignificant, is strongly negative (-0.20). Also, the effect seems to be heterogeneous with respect to positive and negative feedback. While the coefficient of negative feedback is close to zero the coefficient of positive feedback is -0.50 standard deviations and weakly significant. This is a weak indication that revealing the same information at different times may affect beliefs in the effectiveness of effort very differently. While revealing changes in past performance early enough such that one can still react to them by choosing how much to study, revealing the same information in a situation where (test taking) effort is likely already at a maximum because of the high stakes involved, cause students to belief they have less control over their outcomes.

Concerning the positive effects of level feedback in the early treatment, changes in beliefs about the effectiveness do not seem to be a driving factor. Possibly the the effect may be driven by students who learned that their actual rank was lower than their believed rank and who then exerted more learning effort in order to overcome this discrepancy, rather than adjust their relative ability belief downwards.

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 feedback 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 the reference frame of feedback, students in the control group got “good luck” wishes while students in the treatment groups 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 important. Feedback tends to be harmful immediately before the test but significantly increases performance if given 1-3 days in advance. Then it is especially useful for those who recently suffered a decrease in their performance. Furthermore, boys respond more strongly to feedback than girls and students with low self-esteem respond more strongly than students with high-self esteem. We find that competitiveness, confidence in academic abilities, and locus of control do not interact with our feedback intervention. Moreover, we find weak evidence that the change feedback given early had a positive effect on students’ belief that they could affect their outcomes by exerting effort.

Our results give interesting insights into how relative performance feedback works in educational settings and has implications for the design of feedback in all situations where it ability to motivate people is crucial. Moreover, our findings indicate that it may be particularly important to openly give feedback whenever someone decreased their performance but only if they have still have a chance to make up for it.

References

Ashraf, N., Bandiera, O., and Lee, S. S. (2014). Awards unbundled: Evidence from a natural field experiment. *Journal of Economic Behavior & Organization*, 100:44 – 63.

Azmat, G., Bagues, M., Cabrales, A., and Iriberri, N. (2016). 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.

Azmat, G. and Iriberri, N. (2010). 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.

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

Bandiera, O., Larcinese, V., and Rasul, I. (2015). Blissful ignorance? a natural experiment on the effect of feedback on students' performance. *Labour Economics*, 34:13 – 25. European Association of Labour Economists 26th Annual Conference.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory.* Englewood Cliffs, NJ: Prentice Hall.

Barankay, I. (2012). Rank incentives - evidence from a randomized workplace experiment. *unpublished working paper.*

Bell, R. and McCaffrey, D. (2002). Bias Reduction in Standard Errors for Linear and Generalized Linear Models with Multi-stage Samples. *Survey Methodology*, 28:169–179.

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

Blanes i Vidal, J. and Nossol, M. (2011). Tournaments without prizes: Evidence from personnel records. *Management Science*, 57(10):1721–1736.

Bradler, C., Dur, R., Neckermann, S., and Non, A. (2016). Employee Recognition and Performance: A Field Experiment. *Management Science*, accepted.

Buser, T. and Yuan, H. (2016). Do Women give up Competing more easily? Evidence from the Lab and the Dutch Math Olympiad. Discussion Paper TI 2016-096/I, Tinbergen Institute.

Charness, G. and Rabin, M. (2002). Understanding social preferences with simple tests*. *The Quarterly Journal of Economics*, 117(3):817.

Cianci, A. M., Schaubroeck, J. M., and McGill, G. A. (2010). Achievement goals, feedback, and task performance. *Human Performance*, 23(2):131–154.

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

Cutler, D. and Lleras-Muney, A. (2006). Education and Health: Evaluating Theories and Evidence. Working Paper 12352, National Bureau of Economic Research.

Damgaard, M. T. and Nielsen, H. S. (2017). The use of nudges and other behavioural approaches in education. EENEE Analytical Report 29, Prepared for the European Commission.

Dechenaux, E., Kovenock, D., and Sheremeta, R. M. (2015). A survey of experimental research on contests, all-pay auctions and tournaments. *Experimental Economics*, 18(4):609–669.

Deci, E. L., Koestner, R., and Ryan, R. M. (2001). Extrinsic rewards and intrinsic motivation in education: Reconsidered once again. *Review of Educational Research*, 71(1):1–27.

Elsner, B. and Isphording, I. (2017). A big fish in a small pond: Ability rank and human capital investment. *Journal of Labor Economics*, 35(3).

Eriksson, T., Poulsen, A., and Villeval, M. C. (2009). Feedback and incentives: Experimental evidence. *Labour Economics*, 16(6):679–688.

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

Fishbach, A., Eyal, T., and Finkelstein, S. R. (2010). How positive and negative feedback motivate goal pursuit. *Social and Personality Psychology Compass*, 4(8):517–530.

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

Fryer, R., Levitt, S., List, J., and Sadoff, S. (2012). Enhancing the Efficacy of Teacher Incentives through Loss Aversion: A Field Experiment. Working Paper 18237, National Bureau of Economic Research.

Gill, D., Prowse, V., Kissova, Z., and Lee, J. (2016). First-Place Loving and Last-Place Loathing: How Rank in the Distribution of Performance Affects Effort Provision. Discussion Paper 783, Oxford Department of Economics.

Gürtler, O. and Harbring, C. (2010). Feedback in tournaments under commitment problems: Experimental evidence. *Journal of Economics & Management Strategy*, 19(3):771–810.

Hannan, L., Krishnan, R., and Newman, A. (2008). The effects of disseminating relative performance feedback in tournament and individual performance compensation plans. *The Accounting Review*, 83(4):893–913.

Hannan, L., McPhee, G., Newman, A., and Tafkov, I. (2013). The effect of relative performance information on performance and effort allocation in a multi-task environment. *The Accounting Review*, 88(2):553–575.

Hanushek, E. and Rivkin, S. (2006). Teacher Quality. In Hanushek, E. and Welch, F., editors, *Handbook of the Economics of Education*, volume 2, pages 1051–1078. Elsevier.

Hanushek, E., Schwerdt, G., Wiederhold, S., and Wößmann, L. (2015). Returns to Skills around the World: Evidence from {PIAAC}. *European Economic Review*, 73:103–130.

Ilgen, D. R., Fisher, C. D., and Taylor, M. S. (1979). Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology*, 64:349–371.

Jalava, N., Joensen, J. S., and Pellas, E. (2015). Grades and rank: Impacts of non-financial incentives on test performance. *Journal of Economic Behavior & Organization*, 115:161–196.

Kluger, A. N. and DeNisi, A. (1996). 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.

Kluger, A. N. and DeNisi, A. (1998). Feedback Interventions: Toward the Understanding of a Double-Edged Sword. *Current Directions in Psychological Science*, 7(3):pp. 67–72.

Kuziemko, I., Buell, R. W., Reich, T., and Norton, M. I. (2014). ”last-place aversion”: Evidence and redistributive implications. *The Quarterly Journal of Economics*, 129(1):105.

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

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

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

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

Murphy, R. and Weinhardt, F. (2014). Top of the class: The importance of ordinal rank. CESifo Working Paper Series 4815, CESifo Group Munich.

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

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

O'Rourke, E., Haimovitz, K., Ballweber, C., Dweck, C., and Popović, Z. (2014). 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.

Paunesku, D., Walton, G., Romero, C., Smith, E., Yeager, D., and Dweck, C. (2015). Mind-Set Interventions Are a Scalable Treatment for Academic Underachievement. *Psychological Science*, 26(6):784–793.

Rivkin, S., Hanushek, E., and Kain, J. (2005). Teachers, Schools, and Academic Achievement. *Econometrica*, 73(2):417–458.

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

Seligman, M. E. P. (1975). *Helplessness: On depression, development, and death.* CA: Freeman.

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

Tafkov, I. D. (2013). Private and public relative performance information under different compensation contracts. *The Accounting Review*, 88(1):327–350.

Thaler, R., Sunstein, C., and Balz, J. (2013). Chapter 25 - Choice Architecture. In Shafir, E., editor, *The Behavioral Foundations of Public Policy*, pages 428–439. Princeton University Press.

Tran, A. and Zeckhauser, R. (2012). Rank as an inherent incentive: Evidence from a field experiment. *Journal of Public Economics*, 96(9-10):645–650.

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

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

Appendix

A Tables

A.1 Balance and randomization checks

Table 5: 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 6: Randomization Check Class-Level Treatments

	(1)	(2)	(3)	(4)
	Late-Feedback	Early-Feedback	Overall	(1) vs. (2),
	Treatment	Treatment		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 EARLY TIMING and LATE TIMING treatments. Standard errors in parentheses.

Table 7: Randomization Check Student-Level Treatments - 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 Early Timing and Late Timing treatments. Standard errors in parentheses.

Table 8: Randomization Check Student-Level Treatments - Late Timing

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Control	Change	Level	Overall	(1) vs. (2),	(1) vs. (3),	(2) vs. (3),
					p-value	p-value	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 Timing treatment. Standard errors in parentheses.

Table 9: Randomization Check Student-Level Treatments - Early Timing

	(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 Timing treatment. Standard errors in parentheses.

A.2 Raw effects and interaction of timing and reference frame

Table 10: Timing of Feedback - NoControls vs. Controls

	(1)	(2)	(3)	(4)	(5)
Dep. Var: Test Scores	pc_PointsTest3	pc_PointsTest3	pc_PointsTest3	pc_PointsTest3	pc_PointsTest3
Early Timing	0.052	0.050*	0.063*	0.055**	0.056*
	(0.035)	(0.028)	(0.037)	(0.024)	(0.031)
Points Exam 1				0.242***	0.254***
				(0.065)	(0.073)
Points Exam 2				0.389***	0.352***
				(0.079)	(0.081)
Female				−0.022	−0.020
				(0.019)	(0.019)
SchoolFE	No	No	Yes	No	Yes
Pupil Control	No	No	No	Yes	Yes
Class Control	No	Yes	No	No	Yes
N	319	319	319	319	319
adj. R^2	0.018	0.092	0.080	0.367	0.390

Note: This table presents the effect of feedback timing on performance in the last exam using a linear regression model. Dependent variable: percentage points exam 3. Covariates: gender teacher, class size, percentage points exam 1, percentage points exam 2, gender, own room, foreign name, siblings and school fixed effects. Standard errors are reported in parentheses, clustered on classroom level and corrected using biased-reduced linearization. The number of clusters is 19. * p<0.10, ** p<0.05, *** p<0.01.

Table 11: EARLY TIMING vs. LATE TIMING (Interaction with student-level treatments)

Dep. Var: Test Scores	(1) All	(2) If Improved	(3) If Worsened	(4) If Better Half	(5) If Worse Half
Early Timing	0.020	0.015	0.027	0.028	0.025
	(0.031)	(0.055)	(0.037)	(0.049)	(0.048)
Change Frame	−0.005	0.021	−0.038**	0.005	−0.015
	(0.016)	(0.035)	(0.019)	(0.030)	(0.040)
Change Frame × Early Timing	0.040	−0.031	0.106***	0.006	0.071
	(0.028)	(0.057)	(0.034)	(0.039)	(0.059)
Level Frame	−0.025	−0.034	−0.020	−0.048*	−0.004
	(0.018)	(0.039)	(0.042)	(0.025)	(0.039)
Level Frame × Early Timing	0.064**	0.050	0.077	0.080***	0.048
	(0.025)	(0.052)	(0.049)	(0.030)	(0.062)
Points Exam 1	0.255***	0.180	0.511***	0.261***	0.240***
	(0.075)	(0.167)	(0.097)	(0.095)	(0.077)
Points Exam 2	0.353***	0.456***	0.129	0.313**	0.255***
	(0.080)	(0.160)	(0.094)	(0.146)	(0.080)
Female	−0.022	−0.019	−0.015	−0.024	−0.016
	(0.019)	(0.032)	(0.022)	(0.024)	(0.029)
SchoolFE	Yes	Yes	Yes	Yes	Yes
Pupil Control	Yes	Yes	Yes	Yes	Yes
Class Control	Yes	Yes	Yes	Yes	Yes
N	319	163	156	168	151
adj. R^2	0.388	0.282	0.513	0.201	0.313

Note: This table presents the interaction effects of feedback timing and feedback frame using a linear regression model including school fixed effects. Column 1 presents the results for the whole sample, columns 2 and 3 split the sample by students that could improve/worsen their rank from second last to last exam and columns 4 and 5 split the sample by students in the upper/lower half in terms of ranks in the last exam. Dependent variable: percentage points exam 3. Covariates: gender teacher, class size, percentage points exam 1, percentage points exam 2, gender, own room, foreign name, siblings. Standard errors are reported in parentheses, clustered on classroom level and corrected using biased-reduced linearization. The number of clusters is 19. * p<0.10, ** p<0.05, *** p<0.01.

A.3 Sub-group analyses

Table 12: Change Frame vs. Level Frame vs. Control - Class-Level Treatment: Early Timing (Interaction with gender)

	(1)	(2)	(3)	(4)	(5)
Dep. Var: Test Scores	All	If Improved	If Worsened	If Better Half	If Worse Half
Change Frame	0.059**	0.050	0.084**	0.046	0.095**
	(0.024)	(0.054)	(0.041)	(0.051)	(0.037)
Change Frame × Female	-0.051^{*}	-0.119^{*}	-0.006	-0.065	-0.047
	(0.028)	(0.064)	(0.058)	(0.071)	(0.085)
Level Frame	0.074***	0.096***	0.066	0.074*	0.088***
	(0.011)	(0.031)	(0.049)	(0.040)	(0.032)
Level Frame × Female	-0.073^{**}	-0.161^{***}	-0.023	-0.094	-0.064
	(0.031)	(0.053)	(0.074)	(0.063)	(0.069)
Points Exam 1	0.363***	0.326**	0.483***	0.389***	0.397***
	(0.044)	(0.123)	(0.161)	(0.086)	(0.052)
Points Exam 2	0.293***	0.337***	0.148	0.356	0.030
	(0.074)	(0.117)	(0.147)	(0.350)	(0.097)
Female	0.049	0.094*	0.030	0.031	0.071
	(0.038)	(0.054)	(0.055)	(0.048)	(0.073)
ClassFE	Yes	Yes	Yes	Yes	Yes
Pupil Control	Yes	Yes	Yes	Yes	Yes
N	160	87	73	79	81
adj. R^2	0.518	0.444	0.598	0.291	0.549

Note: This table presents the effect of change frame and level frame feedback interacted with students' gender when given 1-3 days in advance using a linear regression model including class fixed effects. Column 1 presents the results for the whole sample, columns 2 and 3 split the sample by students that could improve/worsen their rank from second last to last exam and columns 4 and 5 split the sample by students in the upper/lower half in terms of ranks in the last exam. Dependent variable: percentage points exam 3. Covariates: percentage points exam 1, percentage points exam 2, gender, own room, foreign name, siblings. Standard errors are reported in parentheses, clustered on classroom level and corrected using biased-reduced linearization. The number of clusters is 19. * p<0.10, ** p<0.05, *** p<0.01.

Table 13: Change Frame vs. Level Frame vs. Control - Class-Level Treatment: Late Timing (Interaction with gender)

Dep. Var: Test Scores	(1) All	(2) If Improved	(3) If Worsened	(4) If Better Half	(5) If Worse Half
Change Frame	−0.013	−0.004	−0.030	0.016	0.012
	(0.030)	(0.063)	(0.037)	(0.085)	(0.069)
Change Frame × Female	0.020	0.072	0.006	−0.019	−0.003
	(0.062)	(0.070)	(0.068)	(0.119)	(0.122)
Level Frame	−0.014	−0.041	0.018	−0.017	−0.022
	(0.024)	(0.032)	(0.064)	(0.044)	(0.053)
Level Frame × Female	−0.017	0.086	−0.073	−0.043	0.003
	(0.050)	(0.075)	(0.092)	(0.073)	(0.102)
Points Exam 1	0.122	0.094	0.412^{***}	0.180	0.007
	(0.130)	(0.321)	(0.129)	(0.139)	(0.229)
Points Exam 2	0.435^{***}	0.433	0.227	0.347^{***}	0.476^{***}
	(0.114)	(0.286)	(0.141)	(0.101)	(0.117)
Female	−0.041	-0.104^{**}	0.001	0.001	−0.065
	(0.030)	(0.043)	(0.060)	(0.063)	(0.062)
ClassFE	Yes	Yes	Yes	Yes	Yes
Pupil Control	Yes	Yes	Yes	Yes	Yes
N	159	76	83	89	70
adj. R^2	0.354	0.187	0.451	0.134	0.280

Note: This table presents the effect of change frame and level frame feedback interacted with students' gender when given immediately before the exam using a linear regression model including class fixed effects. Column 1 presents the results for the whole sample, columns 2 and 3 split the sample by students that could improve/worsen their rank from second last to last exam and columns 4 and 5 split the sample by students in the upper/lower half in terms of ranks in the last exam. Dependent variable: percentage points exam 3. Covariates: percentage points exam 1, percentage points exam 2, gender, own room, foreign name, siblings. Standard errors are reported in parentheses, clustered on classroom level and corrected using biased-reduced linearization. The number of clusters is 19. * p<0.10, ** p<0.05, *** p<0.01.

Table 14: Change Frame vs. Level Frame vs. Control - Class-Level Treatment: Early Timing (Interaction with Competitiveness)

	(1)	(2)	(3)	(4)	(5)
Dep. Var: Test Scores	All	If Improved	If Worsened	If Better Half	If Worse Half
Change Frame	0.038	−0.073	0.127**	0.120*	0.039
	(0.041)	(0.067)	(0.054)	(0.060)	(0.089)
Change Frame × High Compet.	−0.005	0.100	−0.096	−0.133*	0.077
	(0.062)	(0.063)	(0.070)	(0.079)	(0.103)
Level Frame	0.047	−0.017	0.122	0.113***	0.073
	(0.054)	(0.074)	(0.077)	(0.041)	(0.110)
Level Frame × High Compet.	−0.020	0.039	−0.119	−0.125**	−0.037
	(0.071)	(0.075)	(0.098)	(0.053)	(0.134)
High Compet.	−0.031	−0.093*	0.045	0.025	−0.017
	(0.043)	(0.049)	(0.056)	(0.046)	(0.110)
Points Exam 1	0.334***	0.288*	0.537***	0.342***	0.393***
	(0.059)	(0.159)	(0.110)	(0.116)	(0.068)
Points Exam 2	0.317***	0.373***	0.111	0.320	−0.007
	(0.065)	(0.132)	(0.113)	(0.286)	(0.135)
Female	−0.002	−0.016	0.015	−0.040	0.029
	(0.026)	(0.055)	(0.021)	(0.033)	(0.026)
ClassFE	Yes	Yes	Yes	Yes	Yes
Pupil Control	Yes	Yes	Yes	Yes	Yes
N	160	87	73	79	81
adj. R^2	0.519	0.425	0.626	0.320	0.556

Note: This table presents the effect of change frame and level frame feedback interacted with competitiveness when given 1-3 days in advance using a linear regression model including class fixed effects. Column 1 presents the results for the whole sample, columns 2 and 3 split the sample by students that could improve/worsen their rank from second last to last exam and columns 4 and 5 split the sample by students in the upper/lower half in terms of ranks in the last exam. Dependent variable: percentage points exam 3. Covariates: percentage points exam 1, percentage points exam 2, gender, own room, foreign name, siblings. Standard errors are reported in parentheses, clustered on classroom level and corrected using biased-reduced linearization. The number of clusters is 19. * p<0.10, ** p<0.05, *** p<0.01.

Table 15: Change Frame vs. Level Frame vs. Control - Class-Level Treatment: Late Timing (Interaction with Competitiveness)

	(1)	(2)	(3)	(4)	(5)
Dep. Var: Test Scores	All	If Improved	If Worsened	If Better Half	If Worse Half
Change Frame	0.036	0.107	−0.016	0.058	0.017
	(0.033)	(0.088)	(0.036)	(0.059)	(0.037)
Change Frame × High Compet.	−0.064	−0.134	−0.031	−0.086	−0.013
	(0.052)	(0.136)	(0.101)	(0.101)	(0.046)
Level Frame	−0.031	−0.059	0.052	−0.035	−0.060
	(0.043)	(0.075)	(0.071)	(0.075)	(0.046)
Level Frame × High Compet.	0.010	0.090	−0.125	−0.005	0.060
	(0.071)	(0.103)	(0.085)	(0.116)	(0.088)
High Compet.	0.011	0.025	0.042	0.015	−0.014
	(0.034)	(0.054)	(0.061)	(0.055)	(0.048)
Points Exam 1	0.113	0.077	0.320**	0.170	−0.007
	(0.118)	(0.257)	(0.123)	(0.131)	(0.155)
Points Exam 2	0.440***	0.424**	0.296**	0.367***	0.464***
	(0.108)	(0.201)	(0.141)	(0.138)	(0.128)
Female	−0.043	−0.035	−0.013	−0.020	−0.074**
	(0.034)	(0.026)	(0.032)	(0.039)	(0.034)
ClassFE	Yes	Yes	Yes	Yes	Yes
Pupil Control	Yes	Yes	Yes	Yes	Yes
N	159	76	83	89	70
adj. R^2	0.357	0.244	0.453	0.136	0.276

Note: This table presents the effect of change frame and level frame feedback interacted with competitiveness when given immediately before the exam using a linear regression model including class fixed effects. Column 1 presents the results for the whole sample, columns 2 and 3 split the sample by students that could improve/worsen their rank from second last to last exam and columns 4 and 5 split the sample by students in the upper/lower half in terms of ranks in the last exam. Dependent variable: percentage points exam 3. Covariates: percentage points exam 1, percentage points exam 2, gender, own room, foreign name, siblings. Standard errors are reported in parentheses, clustered on classroom level and corrected using biased-reduced linearization. The number of clusters is 19. * p<0.10, ** p<0.05, *** p<0.01.

Table 16: Change Frame vs. Level Frame vs. Control - Class-Level Treatment: Early Timing (Interaction with math confidence)

	(1)	(2)	(3)	(4)	(5)
Dep. Var: Test Scores	All	If Improved	If Worsened	If Better Half	If Worse Half
Change Frame	0.052	0.004	0.084**	0.007	0.063
	(0.038)	(0.078)	(0.039)	(0.059)	(0.054)
Change Frame × Confidence	−0.025	0.006	−0.005	0.024	0.008
	(0.062)	(0.096)	(0.074)	(0.100)	(0.082)
Level Frame	0.062*	0.064	0.044*	0.068	0.047
	(0.037)	(0.097)	(0.025)	(0.068)	(0.054)
Level Frame × Confidence	−0.038	−0.048	0.022	−0.050	0.009
	(0.050)	(0.099)	(0.067)	(0.091)	(0.059)
Confidence	0.009	−0.026	0.005	0.013	−0.033
	(0.036)	(0.071)	(0.071)	(0.078)	(0.050)
Points Exam 1	0.363***	0.330**	0.468***	0.374***	0.401***
	(0.047)	(0.136)	(0.125)	(0.087)	(0.048)
Points Exam 2	0.311***	0.398***	0.162	0.380	0.042
	(0.063)	(0.149)	(0.147)	(0.345)	(0.094)
Female	0.006	−0.007	0.021	−0.020	0.033
	(0.027)	(0.053)	(0.023)	(0.044)	(0.029)
ClassFE	Yes	Yes	Yes	Yes	Yes
Pupil Control	Yes	Yes	Yes	Yes	Yes
N	160	87	73	79	81
adj. R^2	0.510	0.413	0.592	0.273	0.542

Note: This table presents the effect of change frame and level frame feedback interacted with confidence when given 1-3 days in advance using a linear regression model including class fixed effects. Column 1 presents the results for the whole sample, columns 2 and 3 split the sample by students that could improve/worsen their rank from second last to last exam and columns 4 and 5 split the sample by students in the upper/lower half in terms of ranks in the last exam. Dependent variable: percentage points exam 3. Covariates: percentage points exam 1, percentage points exam 2, gender, own room, foreign name, siblings. Standard errors are reported in parentheses, clustered on classroom level and corrected using biased-reduced linearization. The number of clusters is 19. * p<0.10, ** p<0.05, *** p<0.01.

Table 16 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 17: Change Frame vs. Level Frame vs. Control - Class-Level Treatment: Early Timing (Interaction with self-esteem)

Dep. Var: Test Scores	(1) All	(2) If Improved	(3) If Worsened	(4) If Better Half	(5) If Worse Half
Change Frame	0.112***	0.096	0.116**	0.009	0.114**
	(0.042)	(0.109)	(0.052)	(0.051)	(0.054)
Change Frame × Self Esteem	−0.118**	−0.134	−0.069	−0.005	−0.106
	(0.049)	(0.104)	(0.069)	(0.073)	(0.079)
Level Frame	0.074**	0.072	0.059	−0.075	0.094***
	(0.031)	(0.096)	(0.037)	(0.054)	(0.032)
Level Frame × Self Esteem	−0.054	−0.066	−0.009	0.149*	−0.106***
	(0.040)	(0.099)	(0.055)	(0.076)	(0.038)
Self Esteem	0.062**	0.050	0.065	−0.083	0.114***
	(0.028)	(0.096)	(0.057)	(0.060)	(0.021)
Points Exam 1	0.360***	0.356**	0.409**	0.437***	0.368***
	(0.040)	(0.146)	(0.159)	(0.111)	(0.068)
Points Exam 2	0.288***	0.330**	0.163	0.287	0.049
	(0.067)	(0.143)	(0.121)	(0.304)	(0.125)
Female	0.014	0.002	0.026	−0.037	0.046
	(0.028)	(0.058)	(0.024)	(0.035)	(0.031)
ClassFE	Yes	Yes	Yes	Yes	Yes
Pupil Control	Yes	Yes	Yes	Yes	Yes
N	160	87	73	79	81
adj. R^2	0.524	0.420	0.611	0.309	0.582

Note: This table presents the effect of change frame and level frame feedback interacted with self esteem when given 1-3 days in advance using a linear regression model including class fixed effects. Column 1 presents the results for the whole sample, columns 2 and 3 split the sample by students that could improve/worsen their rank from second last to last exam and columns 4 and 5 split the sample by students in the upper/lower half in terms of ranks in the last exam. Dependent variable: percentage points exam 3. Covariates: percentage points exam 1, percentage points exam 2, gender, own room, foreign name, siblings. Standard errors are reported in parentheses, clustered on classroom level and corrected using biased-reduced linearization. The number of clusters is 19. * p<0.10, ** p<0.05, *** p<0.01.

A.4 Analysis of effort-effectiveness belief

Table 18: Change Frame vs. Level Frame vs. Control - Class-Level Treatment: Early Timing (Dep. var. effort effectiveness belief)

Dep. Var: Effectiveness Belief	(1) All	(2) If Improved	(3) If Worsened	(4) If Better Half	(5) If Worse Half
Change Frame	0.120^{*}	0.196	0.163	0.269^{*}	−0.070
	(0.066)	(0.154)	(0.130)	(0.142)	(0.137)
Level Frame	0.012	0.103	−0.047	0.068	0.078
	(0.110)	(0.154)	(0.172)	(0.194)	(0.212)
Points Exam 1	0.716^{***}	1.208^{***}	0.658	0.514	0.954
	(0.194)	(0.441)	(1.157)	(0.429)	(0.572)
Points Exam 2	0.909^{**}	0.134	0.899	1.688^{**}	0.723
	(0.399)	(0.782)	(0.934)	(0.736)	(0.675)
Female	−0.056	−0.105	0.045	0.005	−0.049
	(0.084)	(0.073)	(0.187)	(0.086)	(0.190)
ClassFE	Yes	Yes	Yes	Yes	Yes
Pupil Control	Yes	Yes	Yes	Yes	Yes
N	161	88	73	80	81

Note: This table presents the effect of change frame and level frame feedback on effectiveness belief when given 1-3 days in advance using a linear regression model including class fixed effects. Column 1 presents the results for the whole sample, columns 2 and 3 split the sample by students that could improve/worsen their rank from second last to last exam and columns 4 and 5 split the sample by students in the upper/lower half in terms of ranks in the last exam. Dependent variable: percentage points exam 3. Covariates: percentage points exam 1, percentage points exam 2, gender, own room, foreign name, siblings. Standard errors are reported in parentheses, clustered on classroom level and corrected using biased-reduced linearization. The number of clusters is 19. * p<0.10, ** p<0.05, *** p<0.01.

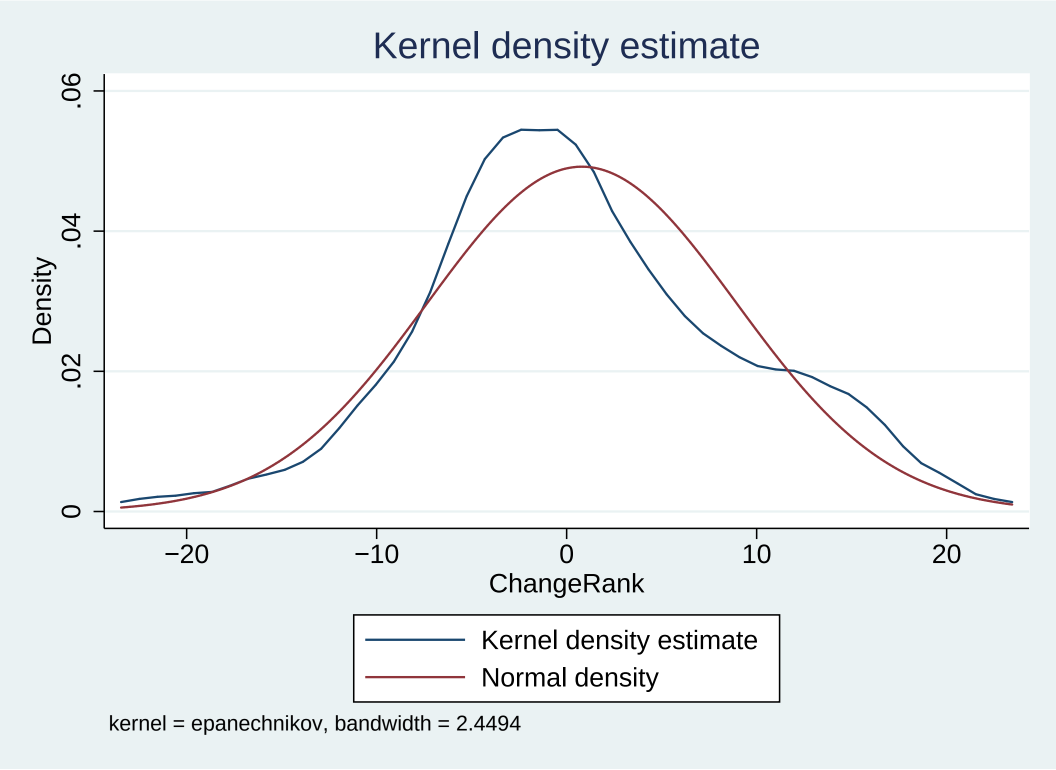
Table 19: Change Frame vs. Level Frame vs. Control - Class-Level Treatment: Late Timing (Dep. var. effort effectiveness belief)

Dep. Var: Effectiveness Belief	(1) All	(2) If Improved	(3) If Worsened	(4) If Better Half	(5) If Worse Half
Change Frame	-0.139	-0.347^{*}	-0.036	-0.210	-0.158
	(0.155)	(0.184)	(0.308)	(0.253)	(0.120)
Level Frame	0.034	-0.455^{**}	0.456^{*}	-0.174	0.277
	(0.152)	(0.184)	(0.258)	(0.265)	(0.187)
Points Exam 1	0.271	0.522	0.611	0.364	0.207
	(0.329)	(0.615)	(0.779)	(0.564)	(0.868)
Points Exam 2	-0.880^{***}	-0.499	-1.734^{**}	-0.475	-1.189
	(0.327)	(0.506)	(0.805)	(0.603)	(1.037)
Female	-0.198	-0.066	-0.228	-0.213	-0.015
	(0.160)	(0.191)	(0.230)	(0.231)	(0.190)
ClassFE	Yes	Yes	Yes	Yes	Yes
Pupil Control	Yes	Yes	Yes	Yes	Yes
N	159	76	83	89	70

Note: This table presents the effect of change frame and level frame feedback on effectiveness belief when given immediately before the exam using a linear regression model including class fixed effects. Column 1 presents the results for the whole sample, columns 2 and 3 split the sample by students that could improve/worsen their rank from second last to last exam and columns 4 and 5 split the sample by students in the upper/lower half in terms of ranks in the last exam. Dependent variable: percentage points exam 3. Covariates: percentage points exam 1, percentage points exam 2, gender, own room, foreign name, siblings. Standard errors are reported in parentheses, clustered on classroom level and corrected using biased-reduced linearization. The number of clusters is 19. * p<0.10, ** p<0.05, *** p<0.01.

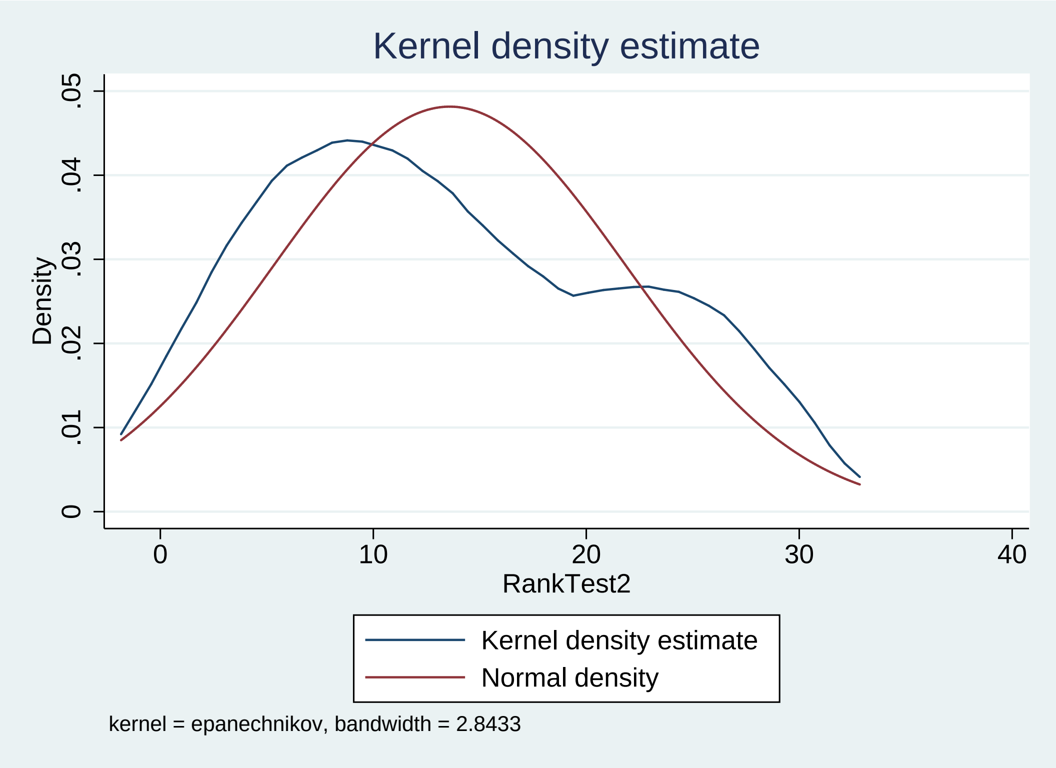
B Graphs

Figure 3: Feedback in Change Frame Treatment



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

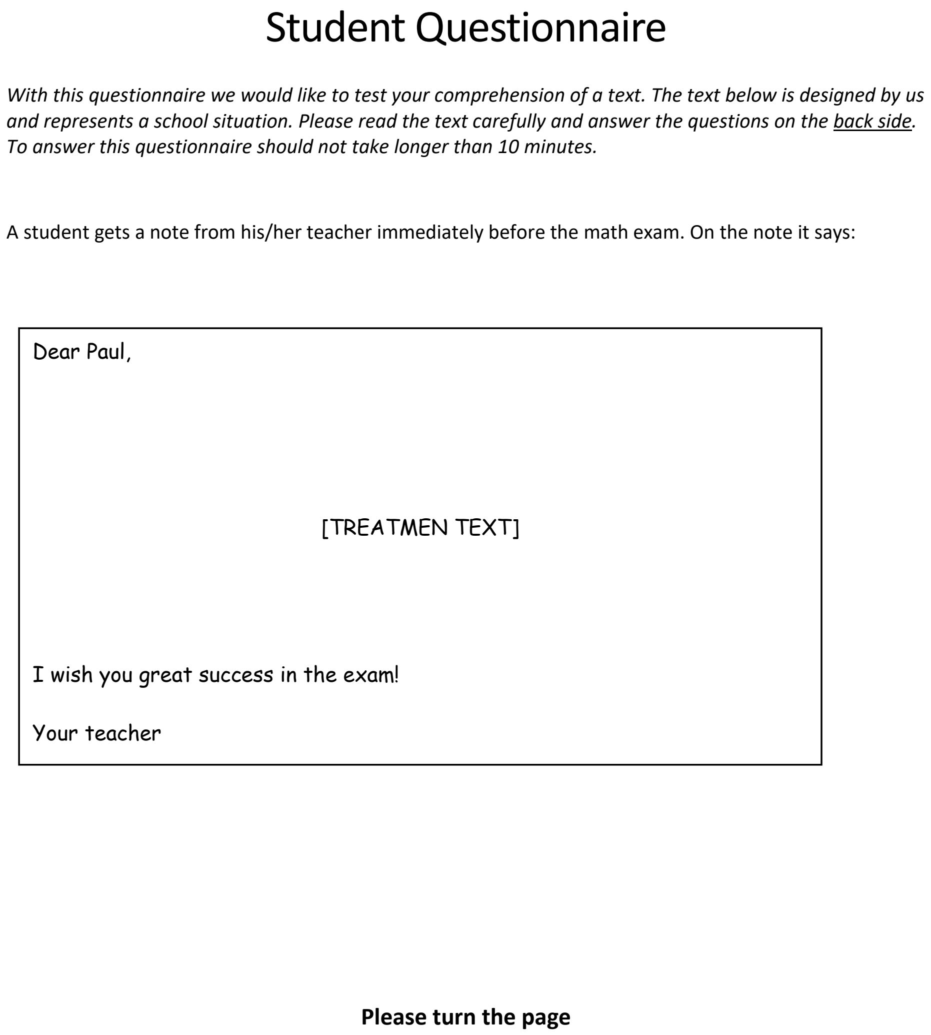
Figure 4: Feedback in Level Frame Treatment



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?

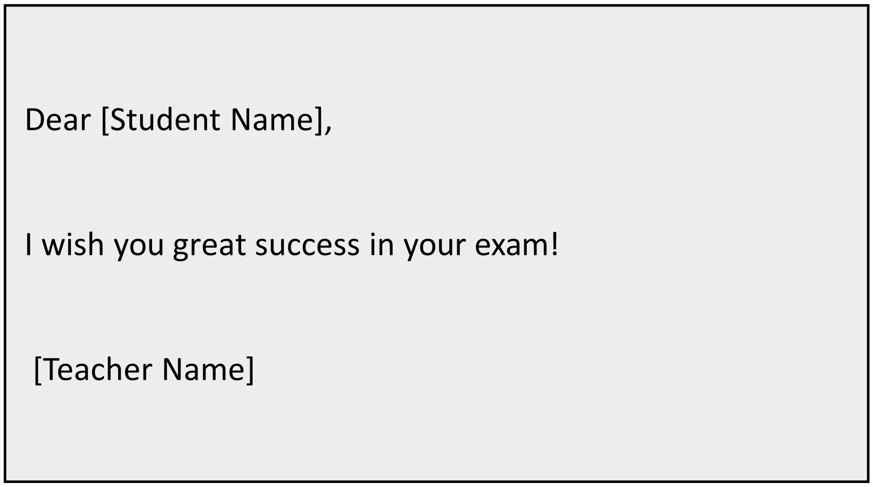
   Number of students in your class:

Thank you very much

D Feedback Notes, Instructions, and Questionnaires

Feedback notes

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

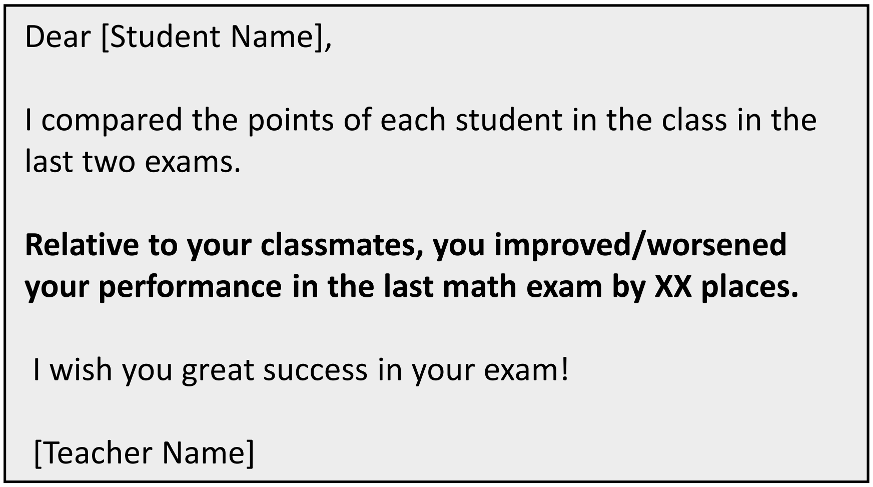


Dear [Student Name],

I wish you great success in your exam!

[Teacher Name]

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



Dear [Student Name],

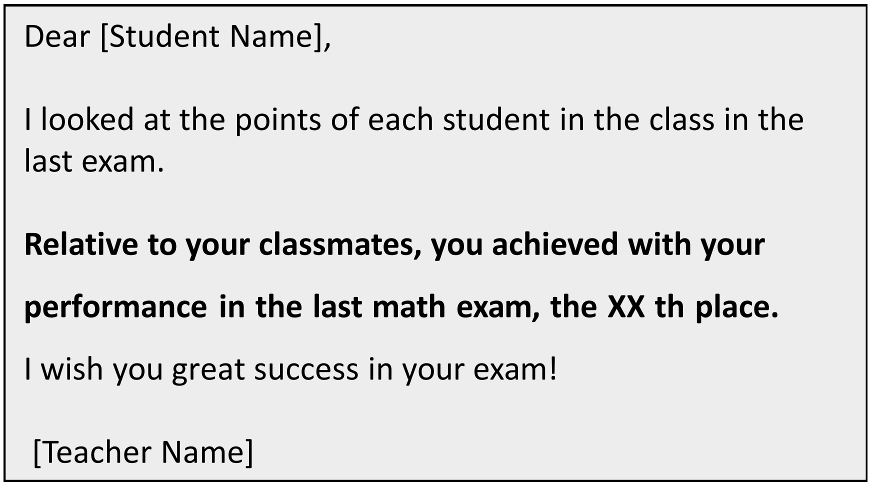
I compared the points of each student in the class in the last two exams.

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 Frame 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.

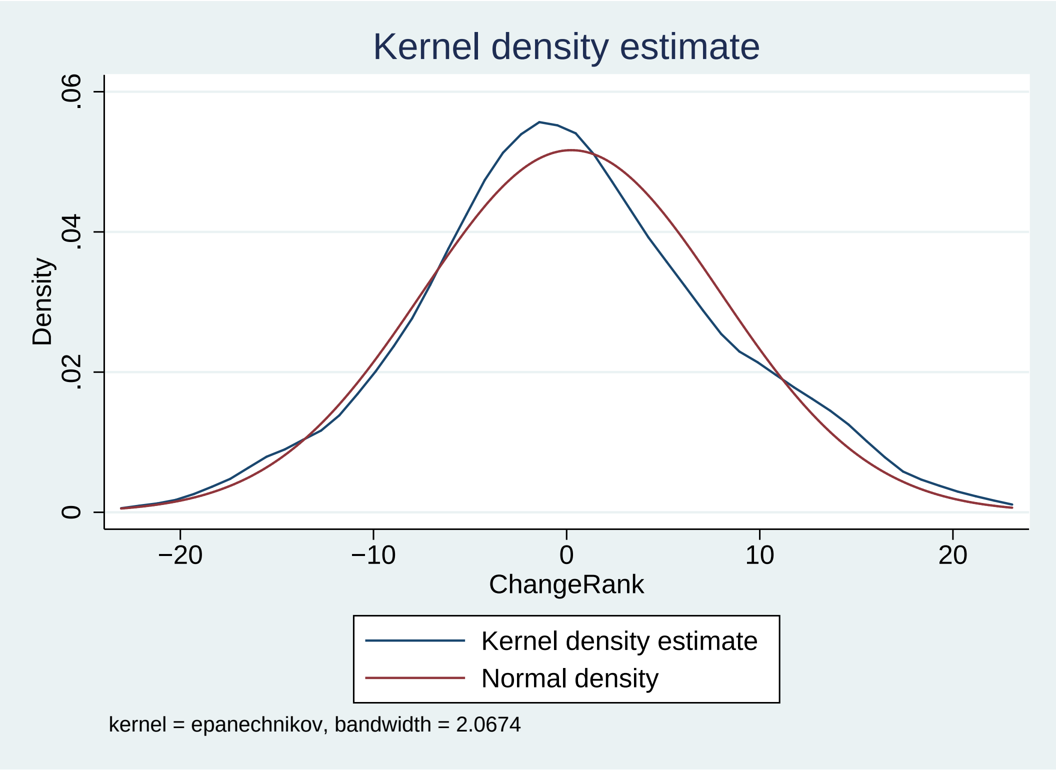
I wish you great success in your exam!

[Teacher Name]

Appendix - Not intended for Publication

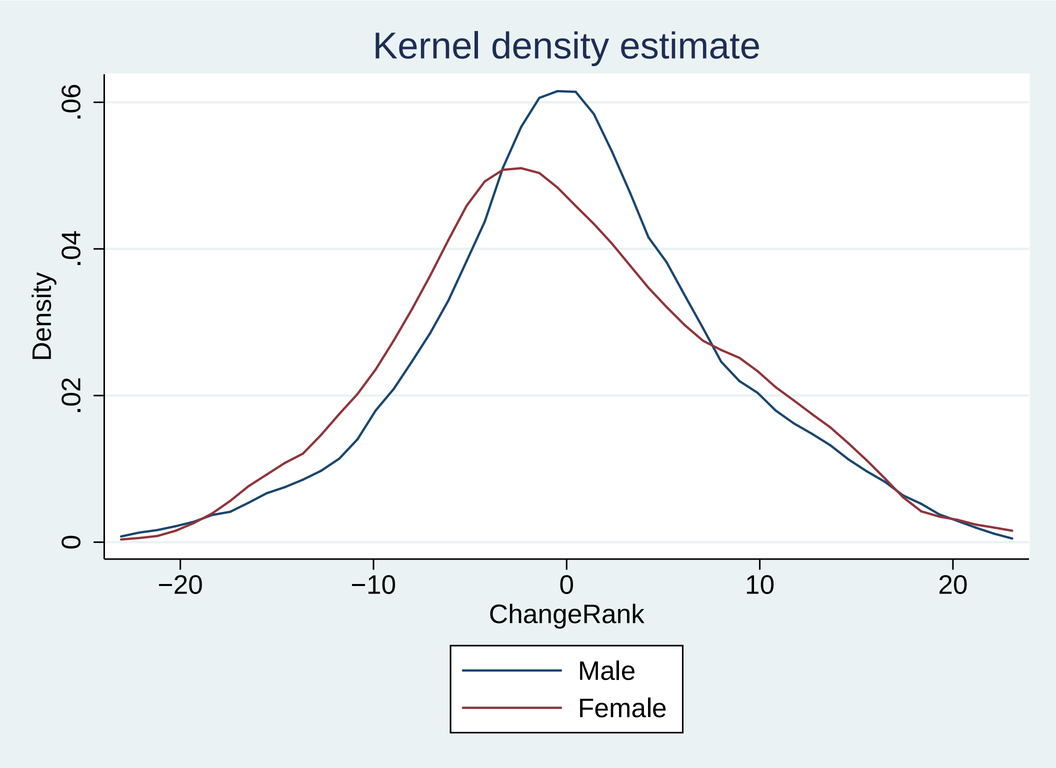
Kernel-density plots

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.