Gutenberg School of Management and Economics
& Research Unit "Interdisciplinary Public Policy"
Discussion Paper Series

Effects of Timing and Reference Frame of Feedback: Evidence from a Field Experiment

Mira Fischer, Valentin Wagner

November 14, 2018

Discussion paper number 1820

Johannes Gutenberg University Mainz
Gutenberg School of Management and Economics
Jakob-Welder-Weg 9
55128 Mainz
Germany
https://wiwi.uni-mainz.de/

Contact details

Mira Fischer
WZB Berlin Social Science Center
Reichpietschufer 50
10785 Berlin
and IZA Institute of Labor Economics
Schaumburg-Lippe-Strasse 5-9
53113 Bonn
Germany

mira.fischer@wzb.eu

Valentin Wagner
Chair of Public and Behavioral Economics
Gutenberg School of Management and Economics
University of Mainz
Jakob-Welder-Weg 4
55128 Mainz
Germany

wagnerv@uni-mainz.de

Effects of Timing and Reference Frame of Feedback: Evidence from a Field Experiment*

Mira Fischer† and Valentin Wagner‡

November 14, 2018

Abstract

Information about past performance has been found to sometimes improve and sometimes worsen subsequent performance. Two factors may help to explain this puzzle: which aspect of one's past performance the information refers to and when it is revealed. In a field experiment in secondary schools, students received information about their absolute rank in the last math exam (level feedback), their change in ranks between the second-last and the last math exam (change feedback), or no feedback. Feedback was given either 1–3 days (early) or immediately (late) before the final math exam of the semester. Both level feedback and change feedback significantly improve students' grades in the final exam when given early and tend to worsen them when given late. The largest effects are found for negative change feedback and are concentrated on male students, who adjust their ability beliefs downwards in response to feedback.

Keywords: timing of feedback, change and level feedback, motivation, field experiment

JEL Codes: I21, M54, D91

*We would like to thank Alexander Cappelen, Thomas Dohmen, Armin Falk, Paul Heidhues, Ingo Isphording, Dorothea Kübler, John List, Henning Müller, Johannes Münster, Gerhard Riener, Matthias Heinz, Hannah Schildberg-Hörisch, Bernd Irlenbusch, Daniel Schunk, Dirk Sliwka, Bertil Tungodden, the participants of the Advances with Field Experiments Conference in Chicago, the Bristol Workshop on Assessment and Feedback, the IZA Workshop on the Economics of Education, the LEER Workshop on Education Economics in Leuven, the Workshop on Education, Skills, and Labor Market Outcomes in Oslo, the Spring Meeting of Young Economists 2018, the International Workshop on Applied Economics of Education in Catanzaro, and the ESA World Meeting 2018, as well as seminar participants in Bergen, Berlin, Bonn, Chicago, Essen, Helsinki, and Maastricht for helpful comments and suggestions. We would like to particularly thank all the teachers and students who participated in the experiment. This research has been conducted with the approval of the ethics committee of the University of Düsseldorf.

†Mira Fischer: WZB Berlin Social Science Center (Reichpietschufer 50, 10785 Berlin, Germany) and IZA Institute of Labor Economics (Schaumburg-Lippe-Strasse 5-9, 53113 Bonn, Germany), mira.fischer@wzb.eu

‡Valentin Wagner: University of Mainz (Jakob-Welder-Weg 4, 55128 Mainz, Germany), wagnerv@uni-mainz.de

1 Introduction

Students and employees are often given feedback about their past performance with the intention to positively influence their future performance. Feedback has indeed sometimes been found to improve performance (Azmat and Iriberri, 2010; Blanes i Vidal and Nossol, 2011; Tran and Zeckhauser, 2012)[1] and may have advantages over monetary incentives as it can be used when the latter are difficult to implement or are not socially accepted. However, feedback is also frequently found to backfire (Barankay, 2012; Ashraf et al., 2014; Azmat et al., 2018; Bradler et al., 2016a) or to be ineffective (Eriksson et al., 2009).[2] Asking which factors are crucial for its success is therefore important.

The influence of a small number of factors on the effectiveness of feedback has already been investigated. For example, the motivational power of feedback has been found to depend on whether a pay-for-performance or a flat rate incentive scheme is present (Azmat and Iriberri, 2016), or whether the information provided is sufficiently precise (Hannan et al., 2008). Furthermore, relative feedback tends to be more effective than performance information referring to an absolute standard (Azmat and Iriberri, 2010). At the same time, there is inconclusive evidence of whether rank information given in public or private is more effective (Tran and Zeckhauser, 2012; Hannan et al., 2013; Tafkov, 2013; Ashraf et al., 2014; Bursztyn and Jensen, 2015; Gill et al., ming).[3] Besides these findings, the question of what makes feedback effective has received rather little attention, leaving unstudied many important aspects that could be relevant for its success as a motivational tool. We begin to fill this gap by asking whether the reference frame—feedback may compare people in terms of their performance *levels* or their performance *changes*—and the timing of relative feedback influence its effectiveness. These seem to be important questions as one would expect both dimensions to matter for people's motivation, and

[1] Economists have investigated different kinds of feedback, such as process feedback (by allowing subjects to observe the behaviors of other people performing the same task, see e.g., Falk and Ichino, 2006; Mas and Moretti, 2009) or outcome feedback (by providing a quantitative measure of past performance such as a test score or rank, see e.g., Tran and Zeckhauser, 2012; Azmat et al., 2018). We will focus on outcome feedback in this study.

[2] See also Kluger and DeNisi (1998) and Hattie and Timperley (2007) for evidence from the psychological literature.

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

insights into their role can potentially be easily incorporated into practice.

In this paper, we study a field experiment in secondary schools in which we exogenously vary *whether* students receive private rank feedback, *when* they receive it, and *what* its standard of comparison (reference frame) is.[4] Within classes, we varied the type of private written feedback students aged around 11–12 years received from their teachers. It either contained information about (i) the absolute rank in the last math exam (level feedback), (ii) the change in ranks between the two previous math exams (change feedback), or (iii) no information (control). Random allocation of students into treatments within classes allows us to control for the heterogeneity of the class environments and to identify the effects of feedback based on comparing students within the same class. Across classes, we varied whether students received their feedback (a) 1–3 days, or (b) immediately, i.e., a few minutes, before the final math exam of the semester. This exam involves high stakes for the students as mathematics is a core subject of the curriculum and students write only three exams per semester in total.[5]

We find (i) the timing of feedback is very important, and (ii) there are similar effects for both feedback types. Feedback given early tends to increase performance while the same feedback given late tends to decrease it. When given a few days before the exam, both change and level feedback significantly increase subsequent performance. In classes with early feedback, students receiving feedback on their rank level significantly increased their performance by 0.2 grade points[6] (3.9 percentage points) compared to students receiving no feedback, while students receiving feedback on rank changes significantly increased their performance by 0.3 grade points (3.8 percentage points). We found it particularly beneficial to inform students of any negative change in performance a few days before the exam as this significantly improved these students' outcomes by 0.6 grade points (8.1 percentage points). In contrast, any feedback given to students immediately before the

[4] A model that motivates the treatment variation (and builds on Fischer and Sliwka, 2018) is presented in the Online Appendix.

[5] Students need an average grade of 3 (on a scale from 1, highest, to 6, lowest) in all subjects for being promoted to the next grade. There are three core subjects—math, German and English. If students get a grade 4 in one core subject, they can only compensate for this grade with a good performance (at least a grade 2) in another core subject.

[6] Grades are given on a 6-point scale.

exam tended to lower subsequent performance. In particular, informing students who had become worse about their negative change in performance immediately before the exam decreased these students' exam grades significantly by 0.3 grade points. Moreover, we find heterogeneous gender effects: While male students respond very strongly to feedback, female students largely do not seem to be affected.

To shed light on the psychological mechanisms that could drive the effects of feedback on performance, we elicit students' confidence in their math ability (belief in competence level in math), their effort-effectiveness belief (perceived relationship between effort and academic outcome) and their state self-esteem (current perception of self-worth and general competence) after the intervention. The analysis shows that boys negatively update their beliefs while girls tend to positively update their beliefs in response to feedback. This may explain why the effects of feedback on performance are largely driven by male students and highlights the importance of information processing and belief elicitation for understanding the effects of feedback on behavior.

To our knowledge, this is the first study identifying the causal effects of the timing of feedback (in terms of interval length to task)[7] and the first to compare the causal effects of two generic types of feedback (level feedback versus change feedback). Our results are relevant for educators because feedback is one of the most important factors related to student achievement (Hattie, 2015), and it is used almost daily by teachers (by means of grades or individual talks).[8] Therefore, our findings—on how to time feedback and which frame of comparison to choose—give guidance for the design of feedback provision in education and, possibly, other settings where the ability to motivate people is crucial.

The paper is organized as follows. The next section gives an overview of the related literature. In section 3 we motivate the treatment variation. Section 4 reports the results of a survey conducted prior to the experiment in which we test whether students of our target age group understand and how they perceive the two types of feedback. Section 5 describes our experimental procedure. Section 6 presents the results and investigates

[7]Psychological studies have e.g., compared immediate feedback and feedback delayed by a few seconds or minutes. See the related literature section for discussion.

[8]See also https://visible-learning.org/hattie-ranking-influences-effect-sizes-learning-achievement/ for a list of factors related to student outcomes.

potential psychological mechanisms. Section 7 concludes.

2 Related Literature

Economists traditionally focus on the introduction of incentives to raise performance but field experiments on monetary and non-monetary (Angrist and Lavy, 2009; Kremer et al., 2009; Fryer, 2013; Bettinger, 2012; Fryer et al., 2012; Levitt et al., 2016) incentives for teachers and/or students in recent years have produced mixed results.

Few studies so far have investigated the effectiveness of feedback to increase performance in the context of education, and those we are aware of have primarily relied on university student samples (exceptions are Azmat and Iriberri, 2010, and Goulas and Megalokonomou, 2015).[9] In an experiment involving Vietnamese university students participating in an English test, Tran and Zeckhauser (2012) provide either private feedback or private plus public feedback on their ranking in in-course mock exams.[10] Overall, the authors find a positive effect of feedback on the final English test and that private plus public feedback tends to outperform private feedback only. This difference, however, was only marginally significant.[11] 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 private, absolute feedback on their past exam performance and others were not. Feedback on exam performance improved future performance mostly for more able students and for students who initially had less information about the academic environment. Azmat et al. (2018) provide 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. In contrast, Brade et al. (2018) give first-year university students in Germany

[9]Damgaard and Nielsen (2018) have recently reviewed the use of feedback and other behaviorally motivated interventions in education.

[10]Private feedback was given by phone and public feedback by postings on the university's noticeboard and website.

[11]In contrast to Tran and Zeckhauser (2012), the results by Ashraf et al. (2014), a study outside the educational context, reveal that private plus public feedback reduces the performance of health workers in Zambia in a nationwide training program.

(normatively framed) relative performance feedback on their accumulated course credits and find an increase in performance only when the feedback is positive.

There is little evidence of the effects of feedback on school-age children (Azmat and Iriberri, 2010; Goulas and Megalokonomou, 2015; Hermes et al., 2018), although schools are a natural setting in which feedback is given almost daily and in which feedback has a potentially large impact on individuals.[12] 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 only informed students of 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. Goulas and Megalokonomou (2015) also exploit data of a natural experiment in high schools in Greece. Similar to Azmat and Iriberri (2010), information about students' performance was made public in some years and—due to a policy reform—private information in other years. In the feedback condition, the students' performance was publicly announced, giving students the opportunity to calculate their national and school rank. In contrast to Azmat and Iriberri (2010), the authors find a positive effect of feedback for the better students. Bursztyn and Jensen (2014) investigate the effect of the announcement of the top three performers in computer-based high school remedial courses on a leader board and find that it decreases overall performance by 13 %. Hermes et al. (2018) study performance transparency in a mathematics e-learning application in a primary school setting. The authors compare a public performance ranking to private individual feedback. While transparency has overall no effects on performance in math, low performers tend to do better and to display higher motivation in the public feedback condition.

[12] In a field experiment in a Dutch school for intermediate vocational education, Buurman et al. (2018) investigate the effect of feedback given by students to teachers on student evaluation scores. They find a zero average treatment effect, but teachers receiving a negative feedback (students' evaluation is lower than teachers' self-assessment) improve significantly in response to receiving feedback.

We are not aware of any studies that investigate the effects of the social reference frame of feedback. The timing of feedback has been investigated with two types of setups. Most studies compared feedback in the form of correct answers *immediately after* subjects completed a task with feedback *delayed* by a few seconds or minutes to test psychological conditioning theory or memory theory (for literature reviews see Smith and Kimball, 2010, and Lechermeier and Fassnacht, 2018). A few studies compare a condition in which feedback is given *immediately after* a prior task with a condition in which feedback is given *immediately before* the subsequent task (Krumhus and Malott, 1980; Bechtel et al., 2015; Henley and Reed, 2015). Our study is the first to compare the effects of giving feedback *days before* the task and *immediately before* the task and hence to investigate the timing of feedback in terms of interval length to the task.

3 Motivation of Treatments

Whether feedback is given a few days before or immediately before a task is potentially crucial for its effects because educational outcomes may be influenced by effort exerted at different times—in preparation and for the task itself (cf. Levitt et al., 2016; Wagner, 2016, for changes in effort in the task itself). While earlier feedback may influence preparation effort, possibly by countering students' tendency to procrastinate and to start preparations too late (Steel, 2007), feedback given more immediately before a task may potentially have a stronger effect on effort in the task itself due to people's tendency to place a greater weight on more recent information (Hogarth and Einhorn, 1992). Furthermore, the timing of the feedback may also matter if it influences both performance expectations and emotions (Loewenstein, 2000; Lane et al., 2005; Kräkel, 2008; Bradler et al., 2016b) and the latter might have stronger effects on motivation in the short run than in the long run (Lempert and Phelps, 2014). For example, someone who learns that his past performance was worse than expected may realize that he has to work harder to attain his desired outcome (positive incentive effect). However, having this overconfidence corrected may involve (temporary) negative emotions that decrease the enjoyment of a task or distract from it (Benabou and Tirole, 2016) and may thus decrease performance in the

short run (negative emotional effect).[13]

We chose to provide students with relative performance information (the rank within their class) as people are strongly motivated by it, even in the absence of any tangible benefits (Charness and Rabin, 2002; Azmat and Iriberri, 2010; Kuziemko et al., 2014; Gill et al., ming). In particular, we vary the reference frame and either provide students with information on the *level* of their rank or their *change* in ranks. We expected that feedback about levels influences students empirical beliefs in different ways than feedback about changes. Building on a model by Fischer and Sliwka (2018) two types of beliefs might matter for how much effort a student invests in the exam: (i) confidence in her past level of math performance, and (ii) confidence in the effectiveness of her effort (i.e., her ability to improve her math performance). Assuming that students at different parts of the ability distribution each strive for exam outcomes within their reach, the model predicts that increasing a student's confidence in her past level of math performance decreases the perceived necessity to invest additional effort in exam preparation to reach the desired outcome in the next exam. Furthermore, according to this model, confidence in the ability to improve one's math performance reduces a person's perceived effort costs and, thus, raising it increases effort. Fischer and Sliwka (2018) find evidence that people's effort in a lab experiment responds as predicted by their model.[14] Similarly, the concept of "growth mindset" in the psychological literature (O'Rourke et al., 2014; Paunesku et al., 2015, which is closely related to the concept of "grit," recently investigated by Alan et al., 2016) suggests that promoting the belief that skills are malleable motivates students to invest more effort in education. We expect that feedback that makes changes in past performance salient strengthens this belief (and reduces perceived effort costs). We compare change feedback to level feedback as level feedback in the form of grades is the standard in educational settings. There is also evidence from the literature that feedback comparing performance levels is often positively motivating, as it may help to correct people's overconfidence with respect to their performance level (Krueger and Mueller,

[13]The importance of timing is also supported by the dual-process theory (Loewenstein, 2000; Alos-Ferrer and Strack, 2014): People's immediate "hot state" response to information differs from their longer-term "cold state" response.

[14]In the online appendix, we present an application of the model to our setting.

2002; Hoelzl and Rustichini, 2005; Malmendier and Tate, 2005; Park and Santos-Pinto, 2010), and thus makes them less confident of having already done "enough" (Azmat et al., 2018; Fischer and Sliwka, 2018), which raises incentives to exert effort.

The effect of feedback on one's level of past performance should depend on whether a person ex-ante is overconfident or underconfident with respect to her level of past performance. If she is overconfident, learning about the true level of past performance is disappointing and will thus lower her confidence in her level of performance (and increase the perceived necessity of effort). If she is underconfident, learning the same information will be positively surprising and will raise her confidence in her level of performance (and decrease the perceived necessity of effort).

4 Pre-test of Treatments

The students in our sample are quite young, and in order to test whether they understand our feedback (to disentangle lack of understanding and ineffectiveness of the information) and how they interpret it (to enable us to interpret possible treatment effects), we conducted a survey in six classes in four schools with a total of 151 students of the same age group as our experimental sample before implementing the field experiment. These children did not participate in the main experiment.

The survey consisted of a two-page questionnaire. On the front of the page students saw a feedback note (of the same types we later used in the experiment) addressing a fictitious student named "Paul" and were asked to imagine themselves in his position. The feedback note contained either the level or the change feedback, and both of them were varied (good ranks to bad ranks, positive and negative change in ranks). On the back of the page, students had to breifly summarize the information on the front of the page and answered a quiz to test whether they had understood it correctly. They were also asked to give their guess of how Paul feels ("very good" to "very bad") after having read the feedback note and of how motivated ("not at all" to "very strongly") Paul would be to exert effort in the next exam. We also asked students whether they knew the number of children in their class, which was crucial for correctly interpreting rank feedback.

Most students correctly understood the feedback notes. 86% of the students could correctly calculate by how much Paul's rank changed, and 95% could correctly determine the position of Paul's rank when given level feedback. Moreover, 86% of students knew the exact size of their class. The mean responses to the questions concerning Paul's emotions and motivation are presented in Figure A.1 in Appendix A. Students believe that Paul would be more motivated when receiving change feedback than when receiving level feedback while they do not indicate that the two feedback types would affect emotions differently. Note that the difference in reported motivation between the change feedback and the level feedback may be driven by the presented ranks. Furthermore, students believe that bad feedback (negative change in ranks or rank level below median) makes a student feel worse than good feedback but that the student's motivation to exert effort is quite high (above 3 on a 5-point scale) and approximately the same with negative and positive feedback.

Overall, the results of the pre-experimental survey indicate that most students of our target age group correctly understood the information contained in the two types of feedback, and that they perceived their content as affecting emotions but did not believe that more negative feedback would generally be less motivating than more positive feedback.

5 Timeline of the Experiment

The experiment was conducted in 19 classes (grades 5 and 6) in seven secondary schools in Germany[15] and was approved by the ethics committee of the University of Düsseldorf. In total, 352 students received parental consent (on average, 73.9% per class) and participated in the experiment in May and June 2016. 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.[16] To train teachers how to conduct the experiment, we visited the schools in the run-up to the experiment. Dur-

[15]Schools are located in the cities of Bonn, Cologne, and Düsseldorf.

[16]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.

ing this meeting, the intervention was explained and teachers' questions were answered. We sent teachers two envelopes with the material needed to run the experiment. A first envelope contained written instructions for the teachers, outlining the time schedule and steps of the intervention, consent forms to be signed by parents and templates (to be returned to researchers) for the test results of the first and the second math exams of the semester (grades and points obtained in each exam and the maximum number of points reachable). Teachers provided us with names, enabling 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. A second envelope was sent to schools a few days before the third exam. It contained the personalized feedback notes, which were sheets of paper that were folded and had the name of the student it referred to clearly written on its outside. The envelope also contained a result template for the third exam and student questionnaires.

Treatment Intervention

We want to test how relative performance feedback affects a student's performance in a high-stakes math exam.[17] Based on a 2x3 design, we vary both the *timing* of the feedback and the *reference frame* of feedback independently.

The timing of the feedback was randomized at the 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). This treatment design allowed us to investigate whether the timing of the feedback matters for exam performance. The reference frame of feedback was randomized at the student level. Within the same class, students with parents' permission to participate received personalized written feedback on their rank level in the last math exam (Level Feedback), on their change in rank between

[17]Providing rank feedback seems promising in light of recent findings that a student's rank within their class or cohort affects later achievement independently of underlying ability (Murphy and Weinhardt, 2014; Elsner and Isphording, 2017). 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, indeed, have a higher likelihood of going to college.

the second-last and the last math exam (Change Feedback), 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 the outside. To personalize the feedback, the note addressed the student by their first name and was signed by the teacher. While students in Control received no information about their past performance, in Change Feedback, students received information about their change in rank but no information on their absolute rank levels.[18] Students in Level Feedback were notified of 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. As students had received their grades in the last two exams after the teachers had graded them (i.e., approximately two and four months before the last exams, respectively), the feedback information served as a reminder that contained more detailed information about different aspects of their relative performance and made different aspects of comparison salient.

To shed light on the channels through which feedback might change students' learning and exam performance, students had to answer a questionnaire after reading the feedback notes (in Early Timing) or after completing the exam (in Late Timing).[19] The questionnaire elicited students' confidence in their mathematics ability, their effort-effectiveness belief, and their state self-esteem to allow us to explore mechanisms as well as gender, character traits, and demographic information that enabled us to explore possible heterogeneities in treatment effects (Ilgen et al., 1979; Lam and Schaubroeck, 2000; Noe, 2000; Fedor et al., 2001; Buser and Yuan, 2016). Confidence in math ability was elicited using the German version of the math efficacy scale included in the OECD's Programme for International Student Assessment (PISA) studies (OECD, 2014; based on Bandura, 1986). To elicit students' effort-effectiveness belief, we asked them how much they be-

[18] See Appendix B for an English translations of the exact wording and layout of the notes.

[19] In Early Timing, students filled in the questionnaire immediately after receiving the feedback notes, while in Late Timing students could only fill in the questionnaire after completing the exam. Furthermore, due to time constraints, in Late Timing, the questionnaire was shorter and contained only some of the scales, of which most only consisted of a subset of items as compared to validated versions used in the Early Timing questionnaire and were included for exploratory reasons. Our main analyses of the psychological mechanisms will therefore be based on the validated scales filled in by students immediately after receiving feedback in Early Timing classes.

lieved their exam outcomes could be affected by their effort. Their state self-esteem was measured using the Rosenberg self-esteem scale (Rosenberg, 1965; German version by von Collani and Herzberg, 2003). Character traits were also elicited with validated scales and included (i) locus of control (adapted from PISA [OECD, 2014]; based on Rotter, 1966), (ii) competitiveness (adapted from PISA [OECD 2014]; based on Owens and Barnes, 1992), and (iii) perseverance (adapted from PISA [OECD 2014]; see OECD, 2013).[20]

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.[21] Upon sending the results of the final exam and the filled-out questionnaires, teachers were asked to fill in a short survey.

6 Results

This section presents the results and 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 investigate the effects of timing and then of the reference frame of feedback. Additionally, we explore psychological mechanisms by which feedback affects outcomes.

6.1 Randomization

Blocked on grade level, classes were randomized into either the Early Timing treatment or the Late Timing treatment. With respect to these class-level treatments, non-random self-selection was possible as parents learned whether feedback would be given 1–3 days before the exam or immediately before the exam. This was necessary so as to receive parents' fully informed consent. However, as we will show, we do not find evidence of strategic self-selection into class-level treatments. Within classes, students were then ran-

[20] For the measures adapted from the PISA studies, also see Marsh et al. (2006).

[21] This was to prevent the feedback notes from being shown to other students (with Early Timing) and from teachers finding them in the exam booklets when they graded the exams (with Late Timing).

domized into the CONTROL group, CHANGE FEEDBACK treatment or LEVEL FEEDBACK treatment. Parents did not learn to which of the three treatments their child was assigned as randomization into student-level treatments took place only after we had obtained parents' consent and students only learned it when they received their feedback notes. Hence, non-random self-selection into the student-level treatments was not possible. We chose to give feedback in the subject of mathematics to minimize the possibility of grade manipulation by teachers. Teachers' discretion in grading is expected to be very small in mathematics compared to a subject where students have to give verbal answers.

Overall, randomization for both class-level and student-level treatments was successful as we find no significant differences between treatments in any relevant dimension (prior test scores and grades, gender, student demographics). Table C.1 in Appendix C reports differences between EARLY TIMING and LATE TIMING. Observables do not differ significantly between these class-level treatments, except with respect to the share of students per class who participated. Surprisingly, the share of participants turned out to be significantly lower in the EARLY TIMING treatment as compared to the LATE TIMING treatment. We expected the opposite as parents might be more concerned about the possible (negative) effects on their children's exam outcomes when feedback was given shortly before the exam.[22] 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, in particular because all relevant characteristics are balanced. Moreover, our analysis controls for teacher grading by accounting for prior test scores and by standardizing test scores at the class level.

Randomization checks for student-level treatments (CHANGE FEEDBACK, LEVEL FEEDBACK, CONTROL) can be found in Table C.2 and Table C.3 in Appendix C. Self-selection into these treatments was not possible, as students had no information on the assignment prior to the intervention, and observables in the student-level treatments are not significantly different from each other.

To summarize, a lower proportion of students participated in the EARLY TIMING

[22]Overall, 26.1% students did not get their parents' consent to participate in the experiment (22.5% in the LATE TIMING treatment and 29.7% in the EARLY TIMING treatment).

treatment. However, student characteristics and prior performance measures do not differ significantly between the class-level and the student-level treatments.

6.2 Data and Descriptive Statistics

Our data consist of pre- and post-intervention performance measures provided by the teachers as well as demographic information and psychological scales from student questionnaires. Importantly, we have detailed information on students' past performance as we know their grades and points in the two last exams before the intervention (written several months earlier) as well as the maximum score possible in these exams.[23] Students were, on average, 11.6 years old with 1.3 siblings. In total, 46.4% of the students were female and 38.0% of students had a non-German first and family name, suggesting the possibility of some migration in the family. Our sample seems to be reasonably representative of secondary school students in the German federal state in which our experimental schools are located (North Rhine-Westphalia), as, overall, 50.8% of secondary school students are female and 41.5% have a migration background.[24] The average grade is 2.74 in exam 1 and 2.59 in exam 2 on a scale from 1 to 6, where 1 is the highest and 6 is the lowest grade.[25] Table 1 presents the number of observations for each treatment cell and summarizes the feedback students received by treatment. It reveals that the range and standard deviation of feedback received in the Change Feedback and Level Feedback treatments are of similar magnitude. Figures D.1 and D.2 in Appendix E show the distribution of feedback pooled over class-level treatments.

[23] See figures H.1 - H.3 in Online Appendix H for the distribution of points in all three exams.

[24] https://www.schulministerium.nrw.de/docs/bp/Ministerium/Service/Schulstatistik/Amtliche-Schuldaten/StatTelegramm2016.pdf

[25] Translation of German grades to American grades: 1.0=A+ or A; 1.3=A-; 1.7=B+, 2.0=B; 2.3=B-; 2.7=C+; 3.0=C; 3.3=C-; 3.7=D+; 4.0=D; >4.0=F (cf. http://german.princeton.edu/wp-content/uploads/2014/11/GPA-Conversion-Chart.pdf)

Table 1: Descriptive statistics of provided feedback

		Obs.	Mean	Std. Dev	Min.	Max.
Change Feedback	Early Timing	59	0.763	8.052	-21	+21
	Late Timing	57	0.842	8.239	-19	+19
Level Feedback	Early Timing	64	13.922	8.407	1	30
	Late Timing	60	13.233	8.208	1	30
Control	Early Timing	55	-	-	-	-
	Late Timing	55	-	-	-	-

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

6.3 Effects of Feedback on Performance

In the following we investigate the effects of our intervention. The following tables present results from linear regressions (OLS) that include prior performance as linear control variables and student characteristics as dummy variables, as well as a constant. Furthermore, regressions analyzing treatments that were randomized at student level contain class fixed effects and regressions analyzing treatments randomized at class level contain school fixed effects. The advantage of including class (school) fixed effects is that we can control for the heterogeneity of the class (school) environments and the identified effects of feedback are based on comparing students within the same class (school). For all presented results, the reported standard errors are clustered at the class level and corrected using bias-reduced linearization (Bell and McCaffrey, 2002; Angrist and Pischke, 2008; Cameron et al., 2008; Cameron and Miller, 2015) to allow for cluster-robust inference with a small number of clusters.

First, we study the effect of *timing* of the feedback on performance to learn whether students who received the intervention 1–3 days before the exam had different outcomes than students who received the intervention immediately before the exam. Then, we look at the Early Timing and the Late Timing groups separately to study the effect of the *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 effects of the Change Feedback, or the Level Feedback or both.

The Role of the Timing of Feedback

We first analyze the effect of feedback on performance in the pooled sample of classes that received the feedback intervention either early or late by comparing students who received any feedback ("Feedback" = 1) with students who received no feedback ("Feedback" = 0). In order to investigate the role of the timing of the feedback, we then interact the treatment dummy "Feedback" with the variable indicating whether the respective class received the intervention early ("EarlyTiming" = 1) or late ("EarlyTiming" = 0). We thus begin by estimating the following OLS model:

$$\begin{aligned} PointsTest3\ (GradeTest3)_i = \\ \alpha + \beta\, Feedback_i + \gamma\, EarlyTiming_j + \delta\, Feedback * EarlyTiming_{ij} \\ + \zeta\, PointsTest1_i + \eta\, PointsTest2_i + \theta\, Covariates_i + \iota\, School_k + \varepsilon_{ijk} \end{aligned} \quad (1)$$

$PointsTest3_i$ 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, $Covariates_i$ is a vector of characteristics of student i: student i's gender, whether student i has a non-German first and family name (to capture migration background), whether student i has siblings, and whether student i has his or her own room at home. $Feedback_i$ indicates whether student i received any type of feedback while $EarlyTiming_j$ indicates that the class was treated 1–3 days before the exam. $School_k$ controls for school fixed effects such that $Feedback_i$ identifies the effect of feedback by comparing the results of students who received feedback with those who did not within the same school.[26] ε_{ijk} is a stochastic i.i.d. error term. While the number of points attained by students in the final exam captures their level of math knowledge, which is the socially relevant outcome, the students themselves might only care about their grade. For this reason, we estimate the model with students' percentage points in the final exam ($PointsTest3_i$) as well as

[26]Note that we have 19 experimental classes in seven schools in total. However, in one school only one class participated and in another school both classes that participated were in the same class-level treatment (this was possible because we blocked randomization at grade level). Thus, in regressions that contain school fixed effects these three classes were dropped, which does not affect any of the main findings.

with the grades attained in this exam ($GradeTest3_i$) as an dependent variable.

Table 2: Effects of Feedback on Performance: The Role of Timing

	Dep. Var.: Points Exam 3		Dep. Var.: Grade Exam 3	
	(1)	(2)	(3)	(4)
Feedback	0.004	-0.021*	-0.006	0.189**
	(0.014)	(0.013)	(0.093)	(0.088)
Early Timing	0.063**	0.027	-0.361*	-0.081
	(0.029)	(0.028)	(0.184)	(0.179)
Feedback * Early Timing		0.053**		-0.409***
		(0.023)		(0.139)
Points Exam 1	0.253***	0.250***	-2.194***	-2.166***
	(0.073)	(0.074)	(0.397)	(0.397)
Points Exam 2	0.373***	0.374***	-2.225***	-2.236***
	(0.088)	(0.088)	(0.483)	(0.478)
Female	-0.023	-0.023	0.100	0.100
	(0.019)	(0.019)	(0.116)	(0.114)
SchoolFE	Yes	Yes	Yes	Yes
Pupil Controls	Yes	Yes	Yes	Yes
N	282	282	282	282
adj. R^2	0.390	0.393	0.421	0.426

Note: This table presents the overall effects of feedback as well as the interaction effects of feedback and timing on performance in the last exam using a linear regression model including school fixed effects. The dependent variable in columns 1 and 2 is percentage points in exam 3. The dependent variable in columns 3 and 4 is the grade in exam 3 (larger grades are worse grades). Covariates: percentage points in exam 1, percentage points in exam 2, gender, own room, foreign name, siblings. Standard errors are reported in parentheses, clustered at class level and corrected using bias-reduced linearization. The number of clusters is 16. * p<0.10, ** p<0.05, *** p<0.01.

The first and third column of Table 2 show that there was no significant overall effect of feedback on points and on grades. However, the second and fourth column show that the effect of feedback strongly depends on its timing. Students who received feedback immediately before the exam had, on average, 2.1 percentage points (0.12 standard deviations) less and about 0.2 worse grades (0.16 standard deviations)—on a scale from 1.0 (best grade) to 6.0 (worst grade)—than students who did not receive any feedback.

These effects are significant at the 10% and the 5% level. Furthermore, when adding up the coefficients of feedback and the interaction term of feedback and timing, we can see that when receiving feedback 1–3 days before the exam students tend to perform better than students who do not receive any feedback. They have an average of 3.2 percentage points (0.18 standard deviations) more and 0.22 points (0.19 standard deviations) better grades than students receiving no feedback. However, the first effect is marginally not significant while the second effect is only significant at the 10% level.[27] These overall effects of timing are interesting because they give a first hint that timing matters for the effects of feedback. Moreover, as we also expect timing to interact with the type of feedback and whether the feedback conveys positive or negative information, we expect the overall effects of timing to disguise a large amount of heterogeneity. For this reason, we will investigate the role of the reference frame of feedback in the next section.[28]

The Role of the Reference Frame of Feedback

In order to investigate the role of the reference frame of feedback, we estimate the following model:

$$PointsTest3_i \ (GradeTest3)_i = \alpha + \beta\, ChangeFeedback_i + \gamma\, LevelFeedback_i + \\ \delta\, PointsTest1_i + \zeta\, PointsTest2_i + \eta\, Covariates_i + \theta\, Class_j + \varepsilon_{ij} \quad (2)$$

$PointsTest3_i$ are the percentage points and $GradeTest3_i$ is the grade of student i in the final math exam. $PointsTest1_i$ and $PointsTest2_i$ are the percentage points in the second-last and the last exam of student i, $Covariates_i$ is the same vector of characteristics of student i as in equation 1. $Class_j$ controls for class fixed effects and ε_{ij} is a stochastic i.i.d. error term.

[27] p=0.112 and p=0.061, respectively, for the combined F-tests of the coefficients of "Feedback" and "Feedback * Early Timing" in columns 2 and 4.

[28] Results when excluding school fixed effects, prior performance measures, and student characteristics can be found in Table E.1 in Appendix E.

Since we found that the timing of the feedback is crucial for its effect, we analyze this model separately for classes that had the intervention 1–3 days before and classes with the intervention immediately before the exam. This allows us to investigate further why students seem to benefit from receiving feedback 1–3 days but not from receiving feedback immediately before the exam.

Effects of change and level feedback when given early Table 3 presents the results with respect to the reference frame of feedback for classes that were treated 1–3 days before the exam. As can be seen in the first and fourth column ("All"), both types of feedback lead to significantly higher exam scores and better grades than those of students in the same classes who did not receive any feedback. Students who received change and students who received level feedback have a 3.8 (0.21 sd) and 3.9 percentage-point (0.22 sd) higher outcome (0.2 and 0.3 points, or 0.19 and 0.22 sd, better grades), respectively, than students in the control group. These effects are significant at the 10% and 5% level (at the 10% and 1% level).

Table 3: Change Feedback vs. Level Feedback vs. Control – Class-Level Treatment: Early Timing

	Dep. Var.: Points Exam 3			Dep. Var.: Grade Exam 3		
	(1) All	(2) Pos Change	(3) Neg Change	(4) All	(5) Pos Change	(6) Neg Change
Change Feedback	0.038*	0.002	0.081***	-0.220*	0.048	-0.588***
	(0.022)	(0.051)	(0.027)	(0.130)	(0.353)	(0.171)
Level Feedback	0.039**	0.026	0.053**	-0.254***	-0.181	-0.386**
	(0.016)	(0.037)	(0.025)	(0.092)	(0.237)	(0.164)
Points Exam 1	0.358***	0.318**	0.473***	-2.581***	-2.367**	-3.655***
	(0.046)	(0.149)	(0.127)	(0.440)	(1.096)	(0.939)
Points Exam 2	0.297***	0.350***	0.161	-1.988***	-2.407**	-0.543
	(0.067)	(0.128)	(0.121)	(0.511)	(0.967)	(0.980)
Female	0.005	-0.006	0.020	-0.017	-0.005	-0.072
	(0.029)	(0.051)	(0.022)	(0.184)	(0.325)	(0.125)
ClassFE	Yes	Yes	Yes	Yes	Yes	Yes
Pupil Controls	Yes	Yes	Yes	Yes	Yes	Yes
N	160	87	73	160	87	73
adj. R^2	0.517	0.426	0.611	0.544	0.481	0.632

Note: This table presents the effect of change feedback and level feedback when given 1–3 days in advance using a linear regression model including class fixed effects. Columns 1 and 4 present the results for the whole sample, columns 2 and 5 present the results for students whose rank improved from exam 1 to exam 2, and columns 3 and 6 present results for students whose rank worsened from exam 1 to exam 2. The dependent variable in columns 1, 2, and 3 is percentage points in exam 3. The dependent variable in columns 4, 5, and 6 is grades in exam 3 (larger grades are worse grades). Covariates: percentage points exam 1, percentage points exam 2, gender, own room, foreign name, siblings. Standard errors are reported in parentheses, clustered at class level and corrected using bias-reduced linearization. The number of clusters is 10. * p<0.10, ** p<0.05, *** p<0.01.

Change feedback differed in its wording depending on whether one experienced a positive or a negative change in ranks. In case of a positive change it said that one's rank "improved," while in case of a negative change it said that one's rank "worsened." For this reason, in columns 2 and 3 we investigate for percentage points outcomes and in columns 5 and 6 examine the grade outcomes to see whether the response to change feedback that reported improvement differs from the response to change feedback that reported worsening. We find that, indeed, it does. Given that one had a positive change in ranks, it does not seem to matter whether students received feedback 1–3 days before

the exam or not as the estimated differences are very small and insignificant. However, if students had a negative change in ranks, receiving this feedback 1–3 days before the exam led students to have a 8.1 percentage-point (0.45 sd) and a 0.6 grade-point (0.50 sd) better outcome than their classmates who did not receive any feedback. Both effects are highly significant at the 1% level.

Interestingly, whether one previously experienced an improvement or worsening in performance also seemed to matter for the feedback on rank level, as the significant effect of level feedback appeared to be driven more strongly by students who had recently suffered a decrease in their performance. It might be that level feedback also tended to be more disappointing, and thus more motivating for students whose performance had dropped.[29] While students in the change feedback treatments most likely interpreted a positive and a negative change feedback as a positive and a negative signal, it is unclear ex ante whether students in the level feedback treatment interpreted their rank feedback as a positive or negative signal. As we elicited students' confidence in their mathematics ability after the intervention, we can investigate the effects of different types of feedback on ability beliefs and will do so further below.

Our results thus provide evidence that, indeed, early change and level feedback significantly improve exam performance.[30]

Effects of change and level feedback when 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, neither change feedback nor level feedback had a significant effect on students' points when they are compared to students within their own class. When looking at students' grades we find a negative effect of 0.3 grade points (0.26 sd) of receiving negative change feedback. This effect is significant at the 5% level. There seems to be heterogeneity in effects. The coefficient of change feedback has a positive sign for students who improved (column 2) and a negative sign for students who worsened

[29] F-tests show that the coefficients of the change feedback and the level feedback in column 3 and column 6, respectively, are not significantly different from each other.

[30] Results when excluding class fixed effects, prior performance measures, and student characteristics can be found in tables E.2 and E.3 in Appendix E.

(column 3). However, these coefficients are not significant.

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

	Dep. Var.: Points Exam 3			Dep. Var.: Grade Exam 3		
	(1) All	(2) Pos Change	(3) Neg Change	(4) All	(5) Pos Change	(6) Neg Change
Change Feedback	-0.002	0.022	-0.029	0.105	-0.037	0.312**
	(0.018)	(0.040)	(0.023)	(0.106)	(0.240)	(0.149)
Level Feedback	-0.022	-0.009	-0.023	0.176	0.025	0.271
	(0.020)	(0.031)	(0.046)	(0.121)	(0.181)	(0.289)
Points Exam 1	0.125	0.105	0.382***	-1.522***	-2.171	-2.832***
	(0.122)	(0.293)	(0.129)	(0.580)	(1.399)	(0.813)
Points Exam 2	0.437***	0.429	0.256*	-2.818***	-2.161	-1.743**
	(0.110)	(0.269)	(0.137)	(0.499)	(1.353)	(0.797)
Female	-0.041	-0.047	-0.021	0.159	0.135	0.093
	(0.031)	(0.039)	(0.028)	(0.160)	(0.162)	(0.205)
ClassFE	Yes	Yes	Yes	Yes	Yes	Yes
Pupil Controls	Yes	Yes	Yes	Yes	Yes	Yes
N	159	76	83	159	76	83
adj. R^2	0.361	0.204	0.456	0.393	0.289	0.436

Note: This table presents the effect of change feedback and level feedback when given immediately before the exam using a linear regression model including class fixed effects. Columns 1 and 4 present the results for the whole sample, columns 2 and 5 present the results for students whose rank improved from exam 1 to exam 2, and columns 3 and 6 present results for students whose rank worsened from exam 1 to exam 2. The dependent variable in columns 1, 2, and 3 is percentage points in exam 3. The dependent variable in columns 4, 5, and 6 is grades in exam 3 (larger grades are worse grades). Covariates: percentage points exam 1, percentage points exam 2, gender, own room, foreign name, siblings. Standard errors are reported in parentheses, clustered at class level and corrected using biased-reduced linearization. The number of clusters is 9. * p<0.10, ** p<0.05, *** p<0.01.

Possible spillover effects when feedback is given early Note that the above analyses identify the effects of feedback on performance by comparing students within Early Timing and Late Timing classes with students who did not receive any feedback. No spillover effects of feedback on students in the control group are possible in late timing classes as students could not find out anything about the feedback other students had received (all students were already seated separately to write the exam and received sheets formatted in the same way). However, the positive effect of feedback in early

TIMING classes could possibly be affected by the spillover effects of our intervention on students who did not receive any feedback. For example, students who found out after our intervention but before the exam that their classmates received feedback while they did not could have been discouraged, leading them to perform worse in the exam compared to a situation where their classmates were not treated. This would cause the positive effects of feedback to be overestimated. Alternatively, the spillover effects could go in the other direction and students who did not receive any feedback in the EARLY TIMING class could, by interacting with those who did receive feedback, become more motivated and perform better in the exam. This would cause us to underestimate the benefits of feedback in EARLY TIMING. To address the question of whether there were spillover effects in EARLY TIMING classes, we compare the results of students in the control groups of EARLY TIMING and LATE TIMING classes. This is a valid procedure because all treatments are balanced in terms of prior performance so potential differences post-intervention performance are due to the intervention.

As can be seen in Table 2 in columns 2 and 4, the control group in classes where spillover effects were possible (EARLY TIMING) tend to have better outcomes (of 2.7 percentage points, or 0.08 grade points) than their counterparts in classes where no spillover effects were possible (LATE TIMING). However, these differences are insignificant. We infer that, if anything, the spillover effects of our intervention on the control group were positive and that possible positive effects of early feedback (level and change feedback) would be lower bound estimates, i.e., we tend to underestimate these treatment effects.

Power calculations We follow Azmat et al. (2018) and assess the power of our analysis in the EARLY TIMING condition and LATE TIMING condition reported in tables 3 and 4 by considering several plausible scenarios with respect to the potential magnitude of the underlying effect. These are (i) the conservative scenario in which the treatment might have no effect or an effect on the outcome variable of 5% of a standard deviation, (ii) the intermediate scenario with an effect size of 15% of a standard deviation, or (iii) the least conservative scenario with an effect size of 25% of a standard deviation. For each class-level treatment (EARLY TIMING and LATE TIMING) we separately calculate

the power of each one of the scenarios, taking the structure of our data into account. In multiple sites, students within a class were randomized into treatments (Control, Change Feedback, Level Feedback). We therefore account for the proportion of explained variance by the blocking variable (the classroom) and the explained variance by the covariates.[31] If the true effect size was 5% of a standard deviation, our field experiment would be able to detect the effect of the treatment on exam performance with a probability of around 12% in both class-level treatments, which would hence be considered under-powered to detect the most conservative scenario. If the true effect size was 15% of a standard deviation, which is an effect size found by Azmat and Iriberri (2010) and Blanes i Vidal and Nossol (2011), we would be able to detect this effect with a probability of around 33% (early) and 27% (late). Our study is able to detect an effect of 25% of a standard deviation with a probability of around 65% (early) and 53% (late). The magnitude of our main findings is in this range and they can thus be identified with sufficient power. It is likely, however, that some of the smaller effects that turn out insignificant in our study would become significant with a larger sample size.

Gender differences While gender differences are not the main focus of our paper, we expect the effectiveness of feedback to depend on ex-ante confidence in the level of past performance and it has widely been shown that male subjects are more confident of their abilities than female subjects (Barber and Odean, 2001; Niederle and Vesterlund, 2007). For this reason, we give a brief report of the results of a heterogeneity analysis by gender with respect to performance. In the next section, we will then investigate the effects of feedback on confidence, which will also allow us to draw inferences with respect to ex-ante gender differences in confidence.

Interestingly, as shown in Table F.1 in Appendix F, the overall positive effect of both change and level feedback in the early treatment is driven by the response of boys. They have a 5.9 and 7.4 percentage-point (0.33 sd and 0.41 sd) better result in the change and level treatment, respectively, than boys 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.

[31] We use the *optimal design* software package for power calculations.

The coefficients of the treatment dummies and the interaction term of the treatment and the female indicator add up to an almost perfect zero effect for both types of feedback. However, we 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 and insignificant. Analyses for classes that received feedback late reveal that neither the results of boys nor the results of girls are influenced by any feedback given late.

6.4 Psychological Mechanisms

In this section we explore several psychological mechanisms that might contribute to explaining the effects of feedback on performance, which has—to our knowledge—not been done in previous studies. First, we look at whether the effects of feedback on outcomes can be explained by changes in students' confidence in their mathematics ability and their belief in the effectiveness of their learning effort. Then, we will investigate whether the feedback influenced students' emotions by analyzing whether it affected their state self-esteem.[32]

Effects of feedback on students' confidence in their mathematics ability As described above, we expect the effectiveness of the feedback to depend on whether the feedback is perceived as positively or negatively surprising. We chose not to elicit students' expectations before the intervention in order to avoid the risk of creating an unnatural framing for the information we would be providing. However, we can indirectly infer whether students were positively or negatively surprised by their feedback by analyzing the effect of feedback on students' confidence elicited after the intervention. If the feedback is positively surprising, students' confidence in their mathematics ability should be positively affected by it relative to the control group, while if the feedback is negatively surprising, students' confidence should be negatively affected by it relative to the control group. As can be seen in Table G.1, level feedback did not have any overall effect on confidence.

[32]Note that, unlike in the regressions with test scores as a dependent variable, we do not have pre-intervention information on students' effort-effectiveness belief or state self-esteem to control for level differences. Effort-effectiveness beliefs and self-esteem were elicited in Early Timing classes as there the feedback intervention was immediately followed by an extended questionnaire.

Interestingly, however, if we split the sample by gender we find that the level feedback significantly decreases boys' confidence by almost 0.3 standard deviations. The effect on girls' confidence is positive and of the same magnitude (but insignificant), which explains why the overall effect of level feedback on confidence is close to zero. The results suggest that boys overestimated their mathematics ability in the absence of feedback and, realizing they were not as good as they thought, they increased learning effort in response to the feedback which then resulted in better performances. Girls, on the other hand, if anything, positively updated their beliefs about their mathematics ability and therefore the feedback did not motivate them to study more. We can also see in Table G.1 that girls generally tend to have lower confidence in their mathematics ability than boys by about 0.3 standard deviations, but this difference is not significant.

Effects of feedback on students' effort-effectiveness belief The "growth mindset" hypothesis described in section 3 predicts that making changes in past performance salient reinforces the belief that one's outcomes can be influenced by one's effort. We therefore expected level feedback not to influence this belief. Table G.2 in Appendix G shows that students who received change feedback report a weakly significant 0.17 standard deviations higher effort-effectiveness belief than students in the control group. Furthermore, the results show that level feedback does not tend to influence this belief overall. When we analyze the effect of feedback on the effort-effectiveness belief separately by gender, we see that, unexpectedly and similar to the effect on confidence in mathematics ability, level feedback has a significantly negative effect on boys while it has a significantly positive effect on girls. It seems that students do not cognitively differentiate between these two ability beliefs to the degree we expected. Rather, positively surprising information, even if it refers to a different ability dimension, seems to raise the effort-effectiveness belief, while negatively surprising information seems to lower it.

Effects of feedback on students' self-esteem We expected that negative change feedback would, on average, be emotionally disappointing to students while positive change feedback would, on average, cheer them up. Furthermore, if students are gen-

erally overconfident, as has been suggested by the literature, the level feedback will be disappointing. Table G.3 in Appendix G shows that feedback tends to have a negative effect on students' self-esteem but the overall estimates are either insignificant or only marginally significant.

As we found that boys largely drive our effects of feedback on performance, and that boys' confidence responds significantly negatively to feedback while girls' confidence tends to respond positively, we expect to see similar patterns with respect to the state self-esteem. In fact, as shown in Table G.3 in Appendix G boys' self-esteem is significantly reduced by about half a standard deviation by both change and level feedback, while girls' self-esteem is increased by change feedback by almost half a standard deviation and does not significantly respond to level feedback. This suggests that the emotional effects of feedback on performance strongly depend on whether the feedback is seen as disappointing or not.

7 Conclusion

We investigate factors that may explain why feedback sometimes has positive and sometimes negative effects on performance. To do so we implemented a randomized feedback intervention in secondary schools. We varied the timing and the reference frame of relative performance feedback to analyze their causal effects on performance in a high-stakes mathematics 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 immediately before the start of the exam. Concerning the reference frame of feedback, students within the same class received either feedback on their absolute rank in the preceding exam, feedback on their change in ranks between the two preceding exams, or no feedback.

We find that level feedback and change feedback significantly improve outcomes in the final exam when given early but tend to decrease outcomes when given shortly before the exam. Moreover, these effects are driven by students who experienced a recent decline in (relative) performance, and feedback has particularly strong effects on boys. We do

not find significant effects of level and change feedback for students who experienced an increase in (relative) performance.

To shed light on potential psychological mechanisms that may contribute toward the explanation of our results, we investigate the effect of feedback on students' confidence in their mathematics ability, their belief in the effectiveness of learning effort, and their state self-esteem. Our results reveal heterogeneous gender effects. While level feedback significantly decreases the state self-esteem, math confidence, and effort-effectiveness belief of the boys, level feedback significantly increased the effort-effectiveness belief of the girls. Moreover, change feedback significantly decreased the state self-esteem of the boys while increasing it for the girls. A straightforward interpretation of these findings is that feedback lowers boys' level of overconfidence which in turn increases their perceived necessity to study for the exam.

Our results suggest that making negative information about past performance salient may significantly improve performance in a high-stakes environment when it is given early enough. However, when it is given too late, a negative emotional effect may dominate a positive incentive effect of information provision. Our results give interesting insights into the psychological and behavioral effects of relative performance feedback in an educational setting—and potentially other situations where the ability to motivate people is crucial—and have two important implications for the design of feedback: (i) feedback works better if given a few days in advance of a high-stakes task, and (ii) teachers should not shy away from giving negative feedback.

References

Alan, S., Boneva, T., and Ertac, S. (2016). Ever failed, try again, succeed better: Results from a randomized educational intervention on grit. *HCEO Working Paper*.

Alos-Ferrer, C. and Strack, F. (2014). From dual processes to multiple selves: Implications for economic behavior. *Journal of Economic Psychology*, 41:1 – 11.

Angrist, J. and Lavy, V. (2009). The effects of high stakes high school achievement awards: Evidence from a randomized trial. *The American Economic Review*, 99(4):1384 – 1414.

Angrist, J. D. and Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press, Princeton NJ.

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. (2018). What you don't know... can't hurt you? A field experiment on relative performance feedback in higher education. *Management Science, forthcoming*.

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.

Bandura, A. (1986). The explanatory and predictive scope of self-efficacy theory. *Journal of Social and Clinical Psychology*, 4(3):359 – 373.

Barankay, I. (2012). Rank incentives – Evidence from a randomized workplace experiment. *Unpublished Working Paper*.

Barber, B. M. and Odean, T. (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *The Quarterly Journal of Economics*, 116(1):261 – 292.

Bechtel, N. T., McGee, H. M., Huitema, B. E., and Dickinson, A. M. (2015). The effects of the temporal placement of feedback on performance. *The Psychological Record*, 65(3):425 – 434.

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.

Benabou, R. and Tirole, J. (2002). Self-confidence and personal motivation. *The Quarterly Journal of Economics*, 117(3):871 – 915.

Benabou, R. and Tirole, J. (2003). Intrinsic and extrinsic motivation. *The Review of Economic Studies*, 70(3):489 – 520.

Benabou, R. and Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives*, 30(3):141 – 164.

Bettinger, E. (2012). Paying to learn: The effect of financial incentives on elementary school test scores. *The 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.

Brade, R., Himmler, O., and Jäckle, R. (2018). Normatively framed relative feedback and performance – Field experiment and replication. Working Paper MPRA Paper No. 88830,.

Bradler, C., Dur, R., Neckermann, S., and Non, A. (2016a). Employee recognition and performance: A field experiment. *Management Science*, 62(11):3085 – 3099.

Bradler, C., Neckermann, S., and Warnke, A. J. (2016b). Incentivizing creativity: A large-scale experiment with tournaments and gifts. *ZEW Discussion Papers*, 16(040).

Bursztyn, L. and Jensen, R. (2014). Should schools recognize or award. *Unpublished Working Paper*.

Bursztyn, L. and Jensen, R. (2015). How does peer pressure affect educational investments? *The Quarterly Journal of Economics*, 130(3):1329 – 1367.

Buser, T. and Yuan, H. (2016). Do women give up competing more easily? Evidence from the lab and the Dutch math olympiad. Working Paper TI 2016-096/I, Tinbergen Institute.

Buurman, M., Delfgaauw, J., Dur, R., and Zoutenbier, R. (2018). The effects of student feedback to teachers: Evidence from a field experiment. Working Paper 18-041/VII, Tinbergen Institute.

Camerer, C. and Lovallo, D. (1999). Overconfidence and excess entry: An experimental approach. *The American Economic Review*, 89(1):306 – 318.

Cameron, A. C. and Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2):317 – 372.

Cameron, C., Gelbach, J., and Miller, D. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414 – 427.

Cameron, J. and Pierce, W. D. (1994). Reinforcement, reward, and intrinsic motivation: A meta-analysis. *Review of Educational research*, 64(3):363 – 423.

Charness, G. and Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, 117(3):817 – 869.

Damgaard, M. T. and Nielsen, H. S. (2018). Nudging in education. *Economics of Education Review*, 64:313 – 342.

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. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125(6):627 – 668.

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):787 – 828.

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.

Fedor, D. B., Davis, W. D., Maslyn, J. M., and Mathieson, K. (2001). Performance improvement efforts in response to negative feedback: The roles of source power and recipient self-esteem. *Journal of Management*, 27(1):79 – 97.

Fischer, M. and Sliwka, D. (2018). Confidence in knowledge or confidence in the ability to learn: An experiment on the causal effects of beliefs on motivation. *Games and Economic Behavior*, 111:122 – 142.

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.

Gervais, S., Heaton, J. B., and Odean, T. (2011). Overconfidence, compensation contracts, and capital budgeting. *Journal of Finance*, 66(5):1735 – 1777.

Gill, D., Kissova, Z., Lee, J., and Prowse, V. (forthcoming). First-place loving and last-place loathing: How rank in the distribution of performance affects effort provision. *Management Science*.

Goulas, S. and Megalokonomou, R. (2015). Knowing who you are: The effect of feedback information on short and long term outcomes. Discussion Paper 1075, University of Warwick, Department of Economics.

Grossman, Z. and Owens, D. (2012). An unlucky feeling: Overconfidence and noisy feedback. *Journal of Economic Behavior & Organization*, 84(2):510 – 524.

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.

Hattie, J. (2015). The applicability of visible learning to higher education. *Scholarship of Teaching and Learning in Psychology*, 1(1):79 – 91.

Hattie, J. and Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1):81–112.

Henley, A. J. and Reed, F. D. D. (2015). Should you order the feedback sandwich? Efficacy of feedback sequence and timing. *Journal of Organizational Behavior Management*, 35(3-4):321 – 335.

Hermes, H., Huschens, M., , Rothlauf, F., and Schunk, D. (2018). Causal effects of relative performance feedback in an e-learning software in primary school. *Unpublished Working Paper*.

Hoelzl, E. and Rustichini, A. (2005). Overconfident: Do you put your money on it? *The Economic Journal*, 115(503):305 – 318.

Hogarth, R. M. and Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, 24(1):1 – 55.

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.

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

Kremer, M., Miguel, E., and Thornton, R. (2009). Incentives to learn. *The Review of Economics and Statistics*, 91(3):437 – 456.

Kräkel, M. (2008). Emotions in tournaments. *Journal of Economic Behavior & Organization*, 67:204 – 214.

Krueger, J. and Mueller, R. A. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology*, 82(2):180 – 188.

Krumhus, K. M. and Malott, R. W. (1980). The effects of modeling and immediate and delayed feedback in staff training. *Journal of Organizational Behavior Management*, 2(4):279 – 293.

Köszegi, B. (2006). Ego utility, overconfidence, and task choice. *Journal of the European Economic Association*, 4(4):673 – 707.

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 – 149.

Lam, S. S. and Schaubroeck, J. (2000). The role of locus of control in reactions to being promoted and to being passed over: A quasi experiment. *Academy of Management Journal*, 43(1):66 – 78.

Lane, A. M., Whyte, G. P., Terry, P. C., and Nevill, A. M. (2005). Mood, self-set goals and examination performance: The moderating effect of depressed mood. *Personality and Individual Differences*, 39:143 – 153.

Lechermeier, J. and Fassnacht, M. (2018). How do performance feedback characteristics influence recipients' reactions? A state-of-the-art review on feedback source, timing, and valence effects. *Management Review Quarterly*, 68(2):145 – 193.

Lempert, K. M. and Phelps, E. A. (2014). Chapter 12 - Neuroeconomics of emotion and decision making. In Glimcher, P. W. and Fehr, E., editors, *Neuroeconomics (Second Edition)*, pages 219 – 236. Academic Press, San Diego.

Levitt, S., List, J., Neckermann, S., and Sadoff, S. (2016). The behavioralist goes to school: Leveraging behavioral economics to improve educational performance. *American Economic Journal: Economic Policy*, 8(4):183 – 219.

Loewenstein, G. (2000). Emotions in economic theory and economic behavior. *The American Economic Review*, 90(2):426 – 432.

Malmendier, U. and Tate, G. (2005). CEO overconfidence and corporate investment. *The Journal of Finance*, 60(6):2661 – 2700.

Marsh, H. W., Hau, K.-T., Artelt, C., Baumert, J., and Peschar, J. L. (2006). OECD's brief self-report measure of educational psychology's most useful affective constructs: Cross-cultural, psychometric comparisons across 25 countries. *International Journal of Testing*, 6(4):311 – 360.

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

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

Niederle, M. and Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *The Quarterly Journal of Economics*, 122(3):1067 – 1101.

Noe, R. A. (2000). Toward an integrative theory of training motivation: A meta-analytic path analysis of 20 years of research. *Journal of Applied Psychology*, 85(5):678 – 707.

OECD (2013). Pisa 2012 results: Ready to learn: Students' engagement, drive and self-beliefs (volume iii).

OECD (2014). Pisa 2012 technical report. `https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf`. Accessed: September 4, 2017.

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.

Owens, L. and Barnes, J. (1992). *Learning preference scales: Handbook and test master set.* Australian Council for Education Research, Victoria.

Park, Y. J. and Santos-Pinto, L. (2010). Overconfidence in tournaments: Evidence from the field. *Theory and Decision*, 69(1):143 – 166.

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.

Rosenberg, M. (1965). *Society and the adolescent self-image. Princeton, NJ: Princeton University Press.* Princeton University Press, Princeton, NJ.

Rotter, J. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied*, 80(1):1 – 28.

Smith, T. A. and Kimball, D. R. (2010). Learning from feedback: Spacing and the delay–retention effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(1):80 – 95.

Steel, P. (2007). The nature of procrastination: A meta-analytic and theoretical review of quintessential self-regulatory failure. *Psychological Bulletin*, 133(1):65 – 94.

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

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.

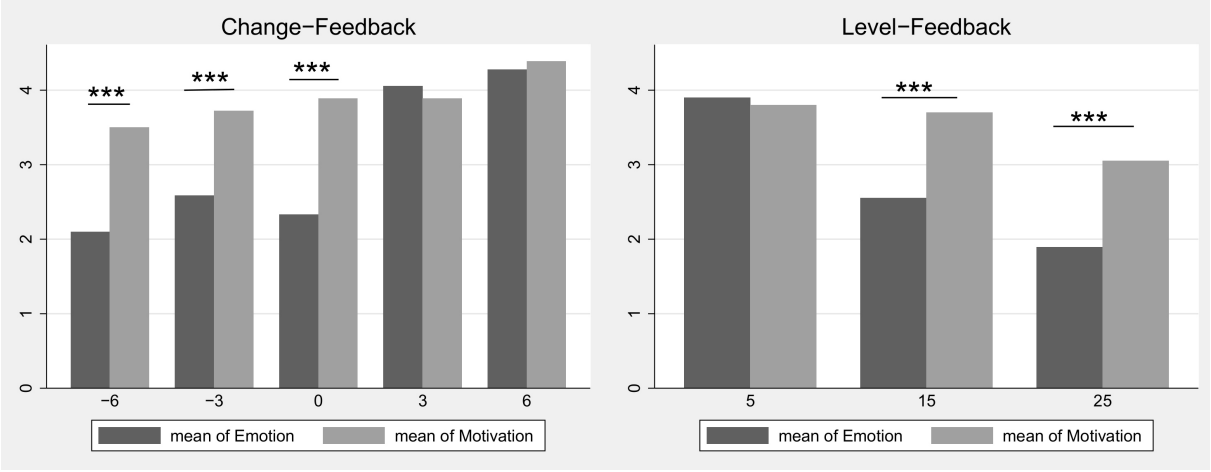
von Collani, G. and Herzberg, P. Y. (2003). Eine revidierte Fassung der deutschsprachigen Skala zum Selbstwertgefühl von Rosenberg. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 24(1):3 – 7.

Wagner, V. (2016). Seeking risk or answering smart? framing in elementary schools. Discussion Paper 227, Düsseldorf Institute for Competition Economics (DICE).

Appendix

A Results of Pre-experimental Survey

Figure A.1: Pretest – Predicted 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 Paul feels after reading the note?* Gray bars are mean responses to the question *How much do you think Paul is 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. The pretest was conducted with 151 students of the same age group (grades 5 and 6) but these students did not participate in the main experiment.

B Feedback Notes

Figure B.1: Feedback Note - Control Group [translated from German]

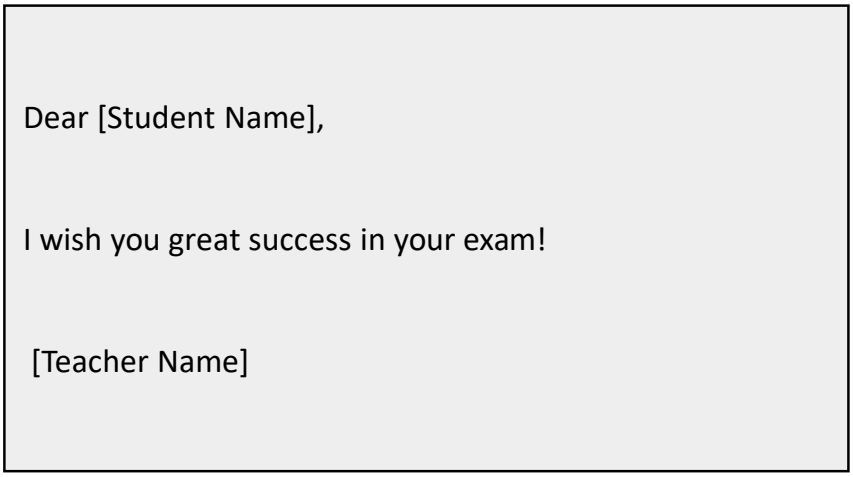


Figure B.2: Feedback Note - Change Feedback Treatment [translated from German]

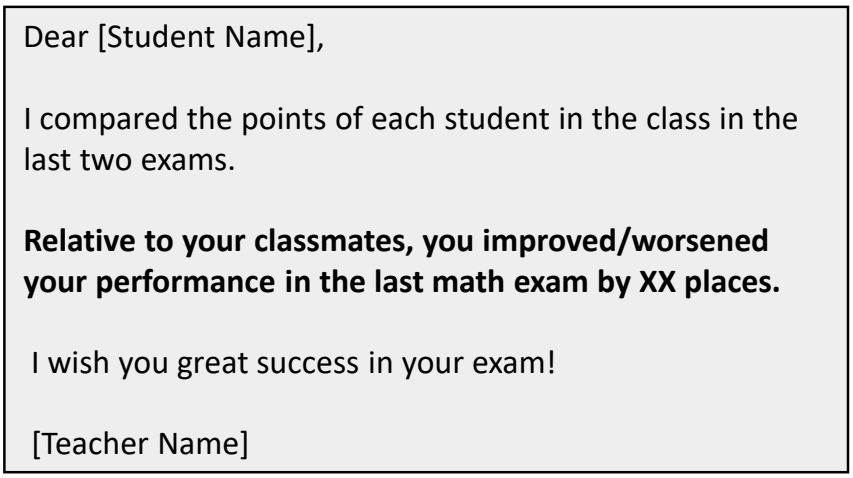


Figure B.3: Feedback Note - Level Feedback Treatment [translated from German]

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 XXth place.

I wish you great success in your exam!

[Teacher Name]

C Balance and Randomization Checks

Table C.1: Randomization Check Class-Level Treatments

	(1) Late-Feedback Treatment	(2) Early-Feedback Treatment	(3) Overall	(4) (1) vs. (2), p-value
Female Teacher	0.793	0.781	0.787	0.781
	(0.031)	(0.031)	(0.022)	
Class Size	27.782	27.242	27.509	0.123
	(0.244)	(0.250)	(0.175)	
Age	23.667	24.708	24.193	0.363
	(0.816)	(0.802)	(0.572)	
Points Exam1	0.712	0.681	0.696	0.105
	(0.014)	(0.014)	(0.010)	
Points Exam2	0.719	0.730	0.725	0.554
	(0.014)	(0.013)	(0.009)	
Rank Exam1	0.495	0.490	0.493	0.889
	(0.022)	(0.021)	(0.015)	
Rank Exam2	0.467	0.493	0.481	0.399
	(0.021)	(0.022)	(0.015)	
Change in Rank	0.523	−0.028	0.243	0.505
	(0.592)	(0.577)	(0.413)	
Share Worsen	0.506	0.455	0.480	0.343
	(0.038)	(0.037)	(0.027)	
Share Participants	0.775	0.703	0.739	0.000
	(0.015)	(0.012)	(0.010)	
Female Pupil	0.480	0.449	0.464	0.570
	(0.038)	(0.037)	(0.027)	
Single Room	0.655	0.596	0.625	0.370
	(0.046)	(0.048)	(0.033)	
Internet	1.115	1.022	1.068	0.366
	(0.072)	(0.073)	(0.051)	
A-Level	2.034	2.056	2.045	0.879
	(0.103)	(0.099)	(0.071)	
Car	1.333	1.303	1.318	0.785
	(0.078)	(0.078)	(0.055)	
Siblings	1.299	1.489	1.395	0.165
	(0.094)	(0.099)	(0.068)	
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 reports group means of key characteristics for the Late Timing (column (1)) and Early Timing (column (2)) treatments. Column (3) presents means for the pooled sample. Variable Age is reported in months starting with children born in June 2002 (Age=1). Standard errors are displayed in parentheses. Column (4) reports the p-values of the two-sided t-test of equality of means between column (1) and column (2).

Table C.2: Randomization Check Student-Level Treatments – Early Timing

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

Note: This table reports group means of key characteristics for the student-level treatments (control, change, level) of Early Timing classes in columns (1)–(3). Variable Age is reported in months starting with children born in June 2002 (Age=1). Standard errors are displayed in parentheses. Columns (4)–(6) report the p-values of the two-sided t-test of equality of means between the treatments.

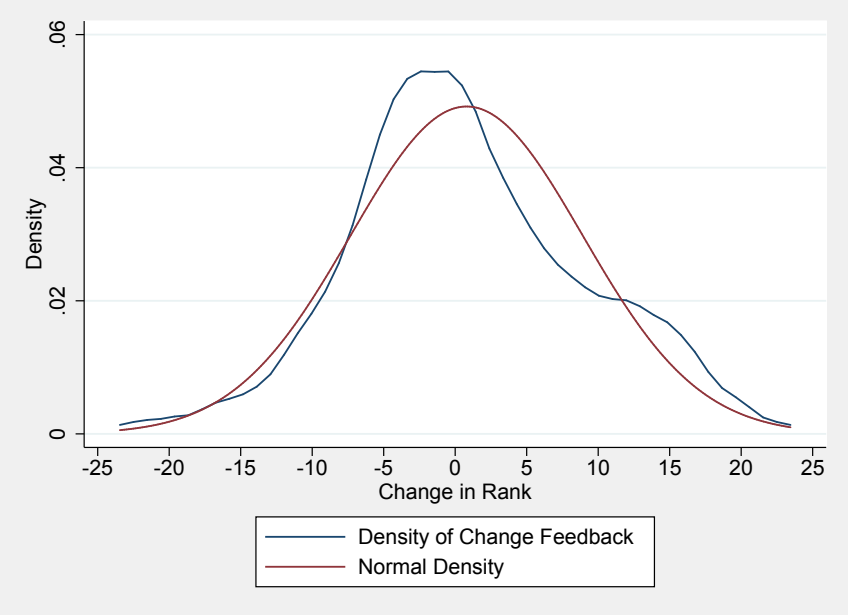
Table C.3: Randomization Check Student-Level Treatments – Late Timing

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

Note: This table reports group means of key characteristics for the student-level treatments (control, change, level) of Late Timing classes in columns (1)–(3). Variable Age is reported in months starting with children born in June 2002 (Age=1). Standard errors are displayed in parentheses. Columns (4)–(6) report the p-values of the two-sided t-test of equality of means between the treatments.

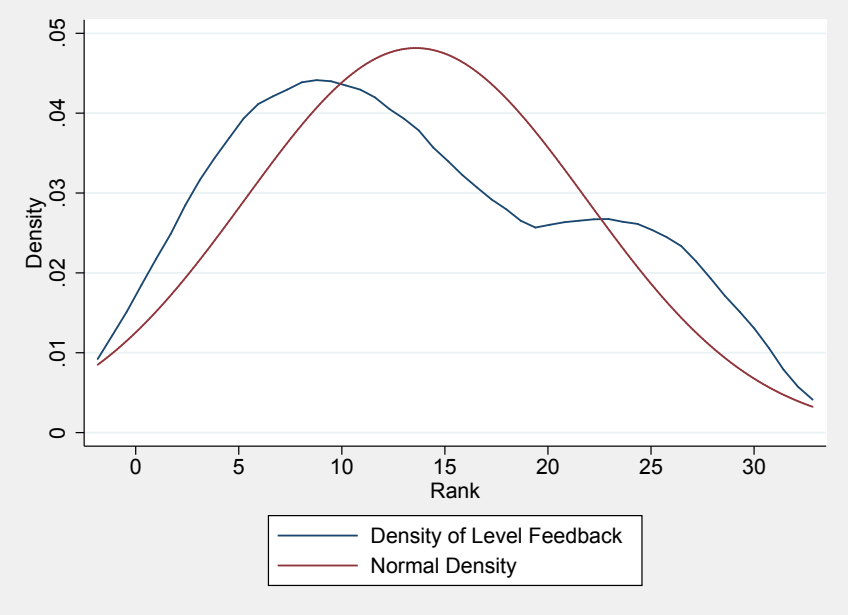
D Graphs

Figure D.1: Feedback in Change Feedback Treatment



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

Figure D.2: Feedback in Level Feedback Treatment



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

E Robustness Checks

Table E.1: Robustness Checks – Class-Level Treatments

	Dep. Var.: Points Exam 3			Dep. Var.: Grade Exam 3		
	(1)	(2)	(3)	(4)	(5)	(6)
Feedback	-0.040**	-0.039**	-0.022	0.326**	0.309**	0.199**
	(0.019)	(0.017)	(0.013)	(0.150)	(0.136)	(0.093)
Early Timing	0.030	0.021	0.032	-0.088	-0.025	-0.129
	(0.035)	(0.040)	(0.026)	(0.207)	(0.233)	(0.168)
Feedback * Early Timing	0.063**	0.062**	0.051**	-0.500***	-0.476**	-0.407***
	(0.026)	(0.025)	(0.023)	(0.193)	(0.187)	(0.141)
Points Exam 1			0.242***			-2.131***
			(0.073)			(0.371)
Points Exam 2			0.374***			-2.229***
			(0.088)			(0.457)
Female			-0.026			0.139
			(0.018)			(0.114)
SchoolFE	No	Yes	No	No	Yes	No
Pupil Controls	No	No	Yes	No	No	Yes
N	282	282	282	282	282	282
adj. R^2	0.041	0.089	0.381	0.034	0.099	0.409

Note: This table presents the effect of feedback timing on performance in the last exam using a linear regression model. The dependent variable in columns 1, 2, and 3 is percentage points in exam 3. The dependent variable in columns 4, 5, and 6 is the grade in exam 3. Columns 1 and 4 do not contain any control variables. Columns 2 and 5 contain school fixed effects but no other control variables. Columns 3 and 6 control for percentage points exam 1, percentage points exam 2, gender, own room, foreign name, and siblings but do not contain school fixed effects. Standard errors are reported in parentheses, clustered at class level and corrected using bias-reduced linearization. The number of clusters is 16. * p<0.10, ** p<0.05, *** p<0.01.

Table E.2: Robustness Checks – Student-Level Treatments in Early Timing – Points

	Dep. Var.: Points in Exam 3								
	All	All	All	Pos Change	Pos Change	Pos Change	Neg Change	Neg Change	Neg Change
Change Feedback	0.027	0.024	0.033	-0.031	-0.033	-0.014	0.095***	0.094***	0.089**
	(0.019)	(0.020)	(0.020)	(0.053)	(0.055)	(0.048)	(0.024)	(0.018)	(0.035)
Level Feedback	0.025	0.029	0.030*	-0.009	0.001	0.016	0.066***	0.073**	0.053
	(0.019)	(0.020)	(0.017)	(0.036)	(0.041)	(0.039)	(0.018)	(0.029)	(0.034)
Points Exam 1			0.314***			0.191			0.511***
			(0.044)			(0.132)			(0.150)
Points Exam 2			0.280***			0.480***			0.081
			(0.106)			(0.084)			(0.187)
Female			-0.006			-0.005			-0.009
			(0.030)			(0.046)			(0.038)
ClassFE	No	Yes	No	No	Yes	No	No	Yes	No
Pupil Controls	No	No	Yes	No	No	Yes	No	No	Yes
N	160	160	160	87	87	87	73	73	73
adj. R^2	-0.008	0.155	0.407	-0.019	0.082	0.363	0.028	0.264	0.482

Note: This table presents the effect of change feedback and level feedback when given 1–3 days in advance using a linear regression model. Dependent variable: percentage points in exam 3. Columns 1, 4, and 7 do not contain any control variables. Columns 2, 5, and 8 contain class fixed effects but no other control variables. Columns 3, 6, and 9 control for percentage points exam 1, percentage points exam 2, gender, own room, foreign name, and siblings but do not contain class fixed effects. Standard errors are reported in parentheses, clustered at class level and corrected using bias-reduced linearization. The number of clusters is 10. * p<0.10, ** p<0.05, *** p<0.01.

Table E.3: Robustness Checks – Student-Level Treatments in Early Timing – Grade

	Dep. Var.: Grade in Exam 3								
	All	All	All	Pos Change	Pos Change	Pos Change	Neg Change	Neg Change	Neg Change
Change Feedback	-0.124	-0.113	-0.177	0.310	0.310	0.173	-0.634***	-0.663***	-0.599**
	(0.136)	(0.131)	(0.120)	(0.378)	(0.377)	(0.337)	(0.213)	(0.124)	(0.268)
Level Feedback	-0.142	-0.172	-0.196*	0.074	-0.000	-0.100	-0.405***	-0.499***	-0.343
	(0.126)	(0.134)	(0.101)	(0.250)	(0.280)	(0.253)	(0.152)	(0.172)	(0.271)
Points Exam 1			-2.214***			-1.524			-3.652***
			(0.439)			(0.991)			(1.096)
Points Exam 2			-1.859**			-3.069***			-0.342
			(0.832)			(0.796)			(1.512)
Female			0.063			0.023			0.099
			(0.201)			(0.289)			(0.262)
ClassFE	No	Yes	No	No	Yes	No	No	Yes	No
Pupil Controls	No	No	Yes	No	No	Yes	No	No	Yes
N	160	160	160	87	87	87	73	73	73
adj. R^2	-0.009	0.153	0.399	-0.012	0.091	0.382	0.025	0.284	0.440

Note: This table presents the effect of change feedback and level feedback when given 1–3 days in advance using a linear regression model. Dependent variable: grade in exam 3. Columns 1, 4, and 7 do not contain any control variables. Columns 2, 5, and 8 contain class fixed effects but no other control variables. Columns 3, 6, and 9 control for percentage points exam 1, percentage points exam 2, gender, own room, foreign name, and siblings but do not contain class fixed effects. Standard errors are reported in parentheses, clustered at class level and corrected using bias-reduced linearization. The number of clusters is 10. * p<0.10, ** p<0.05, *** p<0.01.

Table E.4: Robustness Checks – Student-Level Treatments in Late Timing – Points

	Dep. Var.: Points in Exam 3								
	All	All	All	Pos Change	Pos Change	Pos Change	Neg Change	Neg Change	Neg Change
Change Feedback	-0.018	-0.018	-0.001	-0.028	-0.012	0.015	-0.006	-0.025	-0.011
	(0.023)	(0.021)	(0.017)	(0.030)	(0.031)	(0.037)	(0.044)	(0.030)	(0.025)
Level Feedback	-0.044**	-0.040**	-0.025	-0.035	-0.016	-0.024	-0.059	-0.064*	-0.015
	(0.022)	(0.019)	(0.019)	(0.028)	(0.028)	(0.034)	(0.043)	(0.036)	(0.038)
Points Exam 1			0.180			0.139			0.391***
			(0.114)			(0.258)			(0.126)
Points Exam 2			0.462***			0.435*			0.346***
			(0.107)			(0.248)			(0.116)
Female			-0.044			-0.055			-0.027
			(0.029)			(0.043)			(0.029)
ClassFE	No	Yes	No	No	Yes	No	No	Yes	No
Pupil Controls	No	No	Yes	No	No	Yes	No	No	Yes
N	159	159	159	76	76	76	83	83	83
adj. R^2	-0.001	0.138	0.314	-0.019	0.043	0.183	-0.000	0.242	0.354

Note: This table presents the effect of change feedback and level feedback when given 1–3 days in advance using a linear regression model. Dependent variable: percentage points in exam 3. Columns 1, 4, and 7 do not contain any control variables. Columns 2, 5, and 8 contain class fixed effects but no other control variables. Columns 3, 6, and 9 control for percentage points exam 1, percentage points exam 2, gender, own room, foreign name, and siblings but do not contain class fixed effects. Standard errors are reported in parentheses, clustered at class-level and corrected using bias-reduced linearization. The number of clusters is 9. * p<0.10, ** p<0.05, *** p<0.01.

Table E.5: Robustness Checks – Student-Level Treatments in Late Timing – Grade

	Dep. Var.: Grade in Exam 3								
	All	All	All	Pos Change	Pos Change	Pos Change	Neg Change	Neg Change	Neg Change
Change Feedback	0.213	0.204	0.124	0.247	0.182	-0.022	0.166	0.234	0.266
	(0.177)	(0.156)	(0.105)	(0.222)	(0.237)	(0.209)	(0.339)	(0.240)	(0.177)
Level Feedback	0.347**	0.299**	0.223**	0.116	0.019	0.092	0.614*	0.545*	0.324
	(0.162)	(0.144)	(0.104)	(0.170)	(0.159)	(0.177)	(0.335)	(0.284)	(0.251)
Points Exam 1			-1.860***			-2.196*			-2.888***
			(0.543)			(1.225)			(0.853)
Points Exam 2			-2.652***			-1.923			-1.956***
			(0.462)			(1.164)			(0.630)
Female			0.158			0.165			0.098
			(0.156)			(0.171)			(0.214)
ClassFE	No	Yes	No	No	Yes	No	No	Yes	No
Pupil Controls	No	No	Yes	No	No	Yes	No	No	Yes
N	159	159	159	76	76	76	83	83	83
adj. R^2	0.002	0.113	0.355	-0.019	-0.012	0.282	0.023	0.200	0.375

Note: This table presents the effect of change feedback and level feedback when given 1–3 days in advance using a linear regression model. Dependent variable: grade in exam 3. Columns 1, 4, and 7 do not contain any control variables. Columns 2, 5, and 8 contain class fixed effects but no other control variables. Columns 3, 6, and 9 control for percentage points exam 1, percentage points exam 2, gender, own room, foreign name, and siblings but do not contain class fixed effects. Standard errors are reported in parentheses, clustered at class level and corrected using bias-reduced linearization. The number of clusters is 9. * p<0.10, ** p<0.05, *** p<0.01.

F Gender Differences

Table F.1: Change Feedback vs. Level Feedback vs. Control – Class-Level Treatment: Early Timing (Interaction with gender)

	Dep. Var.: Points in Exam 3			Dep. Var.: Grade in Exam 3		
	All	Pos Change	Neg Change	All	Pos Change	Neg Change
Change Feedback	0.059**	0.050	0.084**	-0.372**	-0.185	-0.688**
	(0.024)	(0.054)	(0.041)	(0.149)	(0.380)	(0.318)
Change Feedback X Female	-0.051*	-0.119*	-0.006	0.351*	0.592*	0.246
	(0.028)	(0.064)	(0.058)	(0.199)	(0.350)	(0.420)
Level Feedback	0.074***	0.096***	0.066	-0.451***	-0.609***	-0.437
	(0.011)	(0.031)	(0.049)	(0.079)	(0.216)	(0.314)
Level Feedback X Female	-0.073**	-0.161***	-0.023	0.417**	0.973***	0.122
	(0.031)	(0.053)	(0.074)	(0.202)	(0.286)	(0.481)
Points Exam 1	0.363***	0.326**	0.483***	-2.605***	-2.430**	-3.701***
	(0.044)	(0.123)	(0.161)	(0.425)	(0.939)	(1.158)
Points Exam 2	0.293***	0.337***	0.148	-1.988***	-2.297***	-0.515
	(0.074)	(0.117)	(0.147)	(0.541)	(0.842)	(1.129)
Female	0.049	0.094*	0.030	-0.289	-0.569*	-0.199
	(0.038)	(0.054)	(0.055)	(0.224)	(0.311)	(0.363)
ClassFE	Yes	Yes	Yes	Yes	Yes	Yes
Pupil Controls	Yes	Yes	Yes	Yes	Yes	Yes
N	160	87	73	160	87	73
adj. R^2	0.518	0.443	0.597	0.543	0.494	0.620

Note: This table presents the effect of change feedback and level feedback interacted with students' gender when given 1–3 days in advance using a linear regression model including class fixed effects. Columns 1 and 4 present the results for the whole sample, columns 2 and 5 present the results for students whose rank improved from exam 1 to exam 2, and columns 3 and 6 present results for students whose rank worsened from exam 1 to exam 2. The dependent variable in columns 1, 2, and 3 is percentage points in exam 3. The dependent variable in columns 4, 5, and 6 is grades in exam 3 (larger grades are worse grades). Covariates: percentage points exam 1, percentage points exam 2, gender, own room, foreign name. Standard errors are reported in parentheses, clustered at classroom level and corrected using biased-reduced linearization. The number of clusters is 10. * p<0.10, ** p<0.05, *** p<0.01.

Table F.2: Change Feedback vs. Level Feedback vs. Control – Class-Level Treatment: Late Timing (Interaction with gender)

	Dep. Var.: Points in Exam 3			Dep. Var.: Grade in Exam 3		
	All	Pos Change	Neg Change	All	Pos Change	Neg Change
Change Feedback	-0.013	-0.004	-0.030	0.210	0.053	0.316
	(0.030)	(0.063)	(0.037)	(0.166)	(0.413)	(0.196)
Change Feedback X Female	0.020	0.072	0.006	-0.208	-0.363	-0.036
	(0.062)	(0.070)	(0.068)	(0.425)	(0.466)	(0.466)
Level Feedback	-0.014	-0.041	0.018	0.186	0.355	0.020
	(0.024)	(0.032)	(0.064)	(0.191)	(0.245)	(0.386)
Level Feedback X Female	-0.017	0.086	-0.073	-0.024	-0.840	0.443
	(0.050)	(0.075)	(0.092)	(0.382)	(0.523)	(0.640)
Points Exam 1	0.122	0.094	0.412***	-1.522***	-2.265	-3.014***
	(0.130)	(0.321)	(0.129)	(0.578)	(1.487)	(0.779)
Points Exam 2	0.435***	0.433	0.227	-2.790***	-2.057	-1.571**
	(0.114)	(0.286)	(0.141)	(0.491)	(1.408)	(0.783)
Female	-0.041	-0.104**	0.001	0.234	0.588**	-0.041
	(0.030)	(0.043)	(0.060)	(0.251)	(0.267)	(0.428)
ClassFE	Yes	Yes	Yes	Yes	Yes	Yes
Pupil Controls	Yes	Yes	Yes	Yes	Yes	Yes
N	159	76	83	159	76	83
adj. R^2	0.353	0.187	0.451	0.386	0.285	0.427

Note: This table presents the effect of change feedback and level feedback interacted with students' gender when given immediately before the exam using a linear regression model including class fixed effects. Columns 1 and 4 present the results for the whole sample, columns 2 and 5 present the results for students whose rank improved from exam 1 to exam 2, and columns 3 and 6 present results for students whose rank worsened from exam 1 to exam 2. The dependent variable in columns 1, 2, and 3 is percentage points in exam 3. The dependent variable in columns 4, 5, and 6 is grades in exam 3 (larger grades are worse grades). Covariates: percentage points exam 1, percentage points exam 2, gender, own room, foreign name, siblings. Standard errors are reported in parentheses, clustered at classroom level and corrected using biased-reduced linearization. The number of clusters is 9. * p<0.10, ** p<0.05, *** p<0.01.

G Psychological Mechanisms: Math Confidence, Effort-effectiveness Belief, and Self-esteem

Table G.1: Change Feedback vs. Level Feedback vs. Control – Dep. var. math confidence

	(1) All	(2) Pos Change	(3) Neg Change	(4) Boys	(5) Girls
Change Feedback	-0.114	-0.308*	0.187	-0.263	-0.210
	(0.184)	(0.182)	(0.172)	(0.222)	(0.372)
Level Feedback	0.048	0.050	0.039	-0.276**	0.314
	(0.116)	(0.155)	(0.261)	(0.129)	(0.260)
Points Exam 1	1.060**	1.402	3.171***	1.428*	0.709
	(0.487)	(1.668)	(0.855)	(0.799)	(0.907)
Points Exam 2	1.285**	0.474	-0.567	0.749	1.826**
	(0.635)	(2.926)	(1.337)	(0.774)	(0.871)
Female	-0.306	-0.271	-0.330		
	(0.196)	(0.185)	(0.429)		
ClassFE	Yes	Yes	Yes	Yes	Yes
Pupil Controls	Yes	Yes	Yes	Yes	Yes
N	154	84	70	83	71
adj. R^2	0.248	0.272	0.223	0.204	0.343

Note: This table presents the effect of change feedback and level feedback on the state self-esteem of students in early treatment classes using a linear regression model including class fixed effects. Column 1 presents results for the whole sample in each early and late treatment classes, column 2 presents results for students who improved, and column 3 presents results for students whose performance worsened from the second-last to the last exam, column 4 presents the results for male students, and column 5 presents the results for female students. Dependent variable: state-self esteem, confidence in mathematics ability (standardized to zero mean and unit standard deviation). Covariates: percentage points exam 1, percentage points exam 2, gender, own room, foreign name, siblings. Standard errors are reported in parentheses, clustered at classroom level and corrected using biased-reduced linearization. The number of clusters is 10. * p<0.10, ** p<0.05, *** p<0.01.

Table G.2: Change Feedback vs. Level Feedback vs. Control – Dep. var. effort-effectiveness belief

	(1) All	(2) Pos Change	(3) Neg Change	(4) Boys	(5) Girls
Change Feedback	0.168*	0.274	0.228	-0.048	0.313
	(0.093)	(0.216)	(0.182)	(0.125)	(0.255)
Level Feedback	0.017	0.144	-0.065	-0.489**	0.478***
	(0.155)	(0.216)	(0.241)	(0.197)	(0.160)
Points Exam 1	1.003***	1.693***	0.922	1.069	0.971
	(0.272)	(0.618)	(1.621)	(0.736)	(0.601)
Points Exam 2	1.273**	0.187	1.259	1.328**	0.953
	(0.559)	(1.095)	(1.309)	(0.626)	(1.000)
Female	-0.079	-0.147	0.063		
	(0.118)	(0.102)	(0.262)		
ClassFE	Yes	Yes	Yes	Yes	Yes
Pupil Controls	Yes	Yes	Yes	Yes	Yes
N	161	88	73	84	77
adj. R^2	0.086	0.119	-0.076	0.189	0.032

Note: This table presents the effect of change feedback and level feedback on the effort-effectiveness belief of students in early treatment classes using a linear regression model including class fixed effects. Column 1 presents results for the whole sample, column 2 presents results for students who improved, column 3 presents results for students whose performance worsened from the second-last to the last exam, column 4 presents the results for male students, and column 5 presents the results for female students. Dependent variable: effort-effectiveness belief (standardized to zero mean and unit standard deviation). Covariates: percentage points exam 1, percentage points exam 2, gender, own room, foreign name, siblings. Standard errors are reported in parentheses, clustered at classroom level and corrected using biased-reduced linearization. The number of clusters is 10. * p<0.10, ** p<0.05, *** p<0.01.

Table G.3: Change Feedback vs. Level Feedback vs. Control – Dep. var. state self-esteem

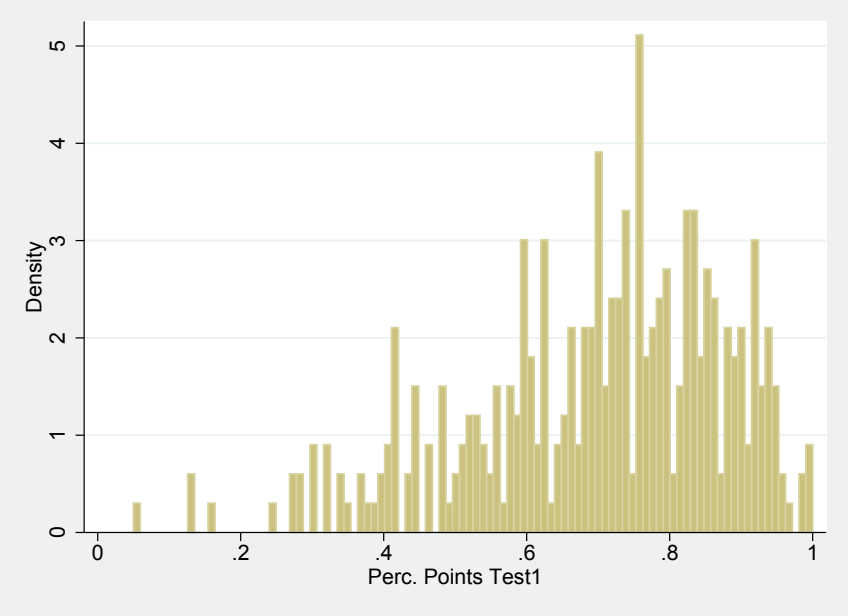
	(1) All	(2) Pos Change	(3) Neg Change	(4) Boys	(5) Girls
Change Feedback	-0.206	-0.437*	-0.075	-0.549**	0.439*
	(0.130)	(0.232)	(0.249)	(0.270)	(0.250)
Level Feedback	-0.280*	-0.442*	-0.058	-0.464***	-0.095
	(0.142)	(0.232)	(0.263)	(0.150)	(0.258)
Points Exam 1	0.715	1.593	0.189	0.867	0.771
	(0.673)	(0.991)	(2.276)	(1.470)	(0.740)
Points Exam 2	1.507***	0.251	1.640	2.041*	0.712
	(0.535)	(1.234)	(1.699)	(1.143)	(0.493)
Female	-0.113	-0.011	-0.028		
	(0.168)	(0.327)	(0.201)		
ClassFE	Yes	Yes	Yes	Yes	Yes
Pupil Controls	Yes	Yes	Yes	Yes	Yes
N	151	81	70	80	71
adj. R^2	0.132	0.147	0.056	0.181	0.119

Note: This table presents the effect of change feedback and level feedback on the state self-esteem of students in early treatment classes using a linear regression model including class fixed effects. Column 1 presents results for the whole sample, column 2 presents results for students who improved, column 3 presents results for students whose performance worsened from the second-last to the last exam, column 4 presents the results for male students, and column 5 presents the results for female students. Dependent variable: state-self esteem (standardized to zero mean and unit standard deviation). Covariates: percentage points exam 1, percentage points exam 2, gender, own room, foreign name, siblings. Standard errors are reported in parentheses, clustered at classroom level and corrected using biased-reduced linearization. The number of clusters is 10. * p<0.10, ** p<0.05, *** p<0.01.

Online Appendix

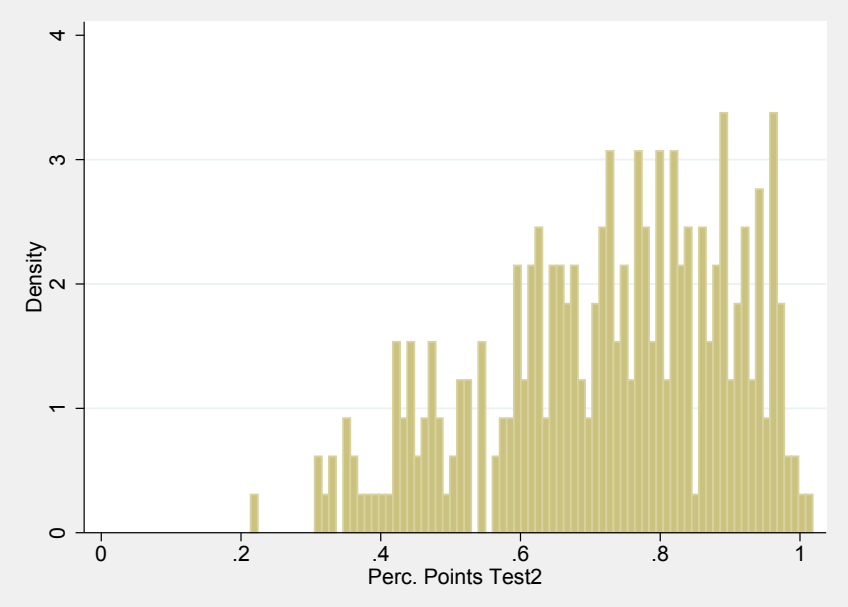
H Distribution of Points

Figure H.1: Distribution of points in Test 1



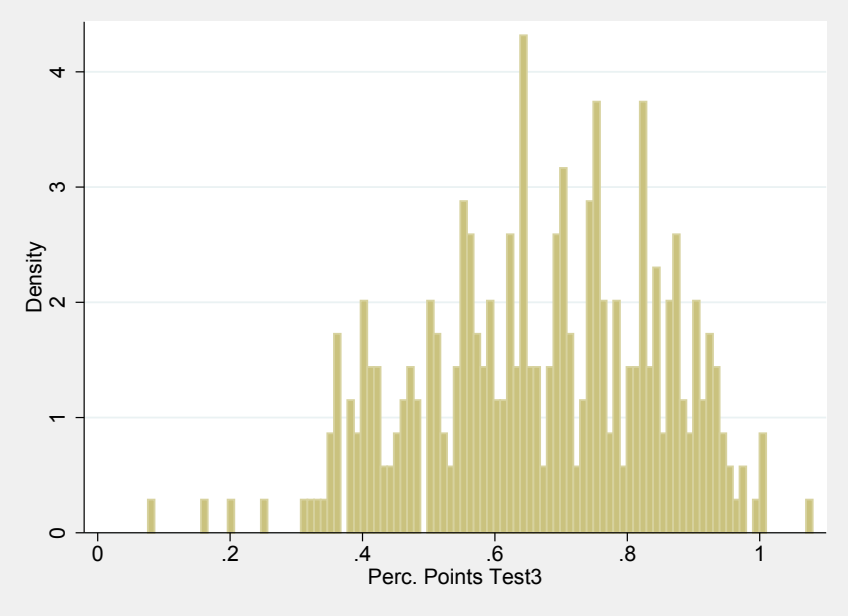
Note: This graph shows the distribution of points in test 1.

Figure H.2: Distribution of points in Test 2



Note: This graph shows the distribution of points in test 2.

Figure H.3: Distribution of points in Test 3



Note: This graph shows the distribution of points in test 3.

I Illustrative Theoretical Framework

We varied both the reference frame and the timing of feedback as we expected that both dimensions might matter for how feedback affects behavior. In the following, we outline why we expected them to matter. We stress, however, that the following exposition is only meant to illustrate possible mechanisms to give guidance on how one could think about the mechanisms of feedback. Thus, we will not derive testable hypotheses, as predictions about how feedback affects an agent's behavior depends on the agent's prior beliefs as well as what information an agent infers from the feedback. We have some evidence for both aspects, but the field setting we study does not allow us to prove any one mechanism. For this reason, the following model should only be understood to motivate the treatment variation and should otherwise be treated as purely illustrative of the mechanisms by which feedback may affect behavior.

Effects of Feedback on ability beliefs and emotions We build on a model by Fischer and Sliwka (2018) and adapt it to our setting in order to illustrate how feedback on levels of performance or feedback on changes in performance may affect effort and in

turn exam outcomes. We assume that a student strives for a certain exam outcome (such as a passing grade or a grade she deems satisfactory) and that how much effort she invests in it depends on (i) confidence in her exam-related prior knowledge, and (ii) confidence in the effectiveness of her effort (i.e., in her academic ability), and (iii) on her emotional state.

A risk-neutral student can invest effort to raise her knowledge. Her knowledge in the test, which determines her exam outcome, is the sum of prior knowledge k and knowledge increases due to (learning and test-taking) effort Δ. Knowledge acquisition is costly and the student's cost function is

$$c(\Delta,\, a,\, p)$$

where a measures the student's effectiveness of effort and p measures her emotional state (the pleasure she experiences while exerting effort). It is assumed that $\frac{\partial c}{\partial \Delta},\ \frac{\partial^2 c}{\partial \Delta^2} > 0$ and $\frac{\partial c}{\partial \Delta \partial a} < 0$ such that the marginal costs of knowledge acquisition are smaller for more academically able students. Furthermore, $\frac{\partial c}{\partial p} < 0$ such that marginal costs of effort are lower if the student experiences more pleasure from exerting effort (cf. Benabou and Tirole, 2002, 2003, 2016; Köszegi, 2006). The student is uncertain about her prior knowledge k and the effectiveness of her effort a . She can receive two types of feedback, the first is a signal of her level of prior knowledge (s_k), while the second is a signal of her effectiveness of effort (s_a), such that $\frac{\partial E[a|s_a,\, s_k]}{\partial s_a} > 0$ and $\frac{\partial E[k|s_a,\, s_k]}{\partial s_k} > 0$. We can decompose $a = E\left[a\,|s_a,\, s_k\right] + \varepsilon_{as}$ and $k = E\left[k\,|s_a, s_k\right] + \varepsilon_{ks}$. A student's emotional state can be influenced temporarily by the signals she receives: Any good news temporarily raises a student's pleasure from exerting effort in a given situation and any bad news decreases it. Psychologists have found that positive competence feedback (even if it does not contain any information, and thus does not influence expectations) may raise intrinsic motivation by raising enjoyment of the task at hand, while negative (non-informational) feedback may have the opposite effect (Cameron and Pierce, 1994; Deci et al., 1999; Fishbach et al.,

2010). Thus, $\frac{\partial E[p|s_a, s_k]}{\partial s_a} > 0$, $\frac{\partial E[p|s_a, s_k]}{\partial s_k} > 0$ and $p = E[p\,|s_a,\, s_k] + \varepsilon_{as} + \varepsilon_{ks} + \varepsilon_{as}\varepsilon_{ks}$ [33] The student has beliefs about both her prior knowledge ($\hat{k}$) and the effectiveness of her effort ($\hat{a}$) which are equal to her expectations about a and k conditional on the signals she receives:

$$\hat{k} = E[k\,|s_a, s_k]$$

$$\hat{a} = E[a\,|s_a, s_k]$$

Furthermore, a student's emotional state is dependent on the two signals: $\hat{p} = E[p\,|s_a,\, s_k]$.

Students attain their desired exam outcome, if $k + \Delta$ exceeds a threshold value τ. In that case they will receive reward B (i.e. a satisfactory exam outcome). Students' objective function can thus be denoted as

$$\max_{\Delta} \Pr\left(\hat{k} + \varepsilon_{ks} + \Delta > \tau\right) B - E[c(\Delta,\, a,\, p)\,|s_a,\, s_k].$$

The first derivative of the objective function is

$$g_{\varepsilon_{ks}}\left(\tau - \hat{k} - \Delta\right) B - E\left[\frac{\partial c(\Delta,\, \hat{a} + \varepsilon_a,\, \hat{p})}{\partial \Delta}\right]$$

where $g_{\varepsilon_{ks}}$ is the density of the error term associated with the signal s_k. Under the condition that the objective function is strictly concave, such that it has a unique solution, it can be shown that:

Proposition *Effort is strictly increasing in the student's confidence in the effectiveness of her effort $\hat{a}$ as well as in her intrinsic motivation $\hat{p}$. It is strictly decreasing in the student's confidence in prior knowledge $\hat{k}$ if and only if $\hat{k}$ is larger than a cut-off value and otherwise strictly increasing.*

The intuition behind this proposition is the following: All students will exert more

[33] Assume that ε_{as} and ε_{ks} are uncorrelated with the signals (s_a, s_k), have mean zero, and unimodal densities with $g'_{\varepsilon_{as}}(0) = g'_{\varepsilon_{ks}}(0) = 0$.

effort when their confidence in their academic ability ($\hat{a}$) or their pleasure from exerting effort ($\hat{p}$) increases as this causes them to perceive their marginal costs of effort to be lower. Benabou and Tirole (2002) call this the "motivation value" of higher confidence. Making students who already perceive their level of prior knowledge as high (relative to their desired outcome) even more confident about it will decrease effort as it makes them more certain that they do not need to invest more effort to achieve their desired exam outcome. However, raising the confidence in prior knowledge ($\hat{k}$) of students who perceive their prior knowledge to be very low will increase effort as they perceive greater chances that their effort will help them to attain their desired outcome.

Effects of feedback on extrinsic and intrinsic motivation

A student's desired outcome likely depends on where in the ability distribution within the class she is. If she is very good at math compared to her classmates, she will likely strive for a top grade (i.e., a top position in the ranking), if she is about average she might strive for an average or somewhat above average grade, while if she knows she is rather weak at math she will just strive for a passing grade (i.e., avoiding the last places in the ranking). Giving students level feedback will disappoint those who overestimate their prior performance while it will positively surprise those who underestimate their prior performance. It has often been found that people tend to overestimate their abilities (Camerer and Lovallo, 1999; Park and Santos-Pinto, 2010; Gervais et al., 2011; Grossman and Owens, 2012).[34] We expect that most students will be disappointed by the level feedback they receive. In the short run this will worsen their emotional state and decrease their intrinsic motivation to exert effort, but they learn that they need to exert more effort than expected in order to attain their desired outcome which raises their extrinsic motivation to exert effort.

[34]Students in our sample, on average, indicate above average confidence: The average student stated that the past two exams were "rather easy" (indicating an average value of 2.2 for both the first and the second test on a scale from 1 to 4, both of which are significantly smaller than 2.5 at the 1%-level (t-test)). The average student also stated they they felt "rather well" prepared for the two past exams (indicating an average value of 2.9 for the first test and 3.0 for the second test on a scale from 1 to 4, both of which are significantly larger than 2.5 at the 1%-level (t-test)). The average student's confidence in math ability is 0.58 on a scale from 0 to 1, where boys have an average value of 0.61 and girls of 0.55. This is a difference of 0.31 standard deviations, which is significant at the 1%-level (t-test). It is a common finding that males are more confident in their abilities than females (see e.g., Barber and Odean, 2001; Niederle and Vesterlund, 2007). All tests in this paper are two-sided.

Change feedback contains information about how a student's performance changed relative to her classmates from the second-last to the last math exam. If we assume that the second-last and the last exam measure the same dimension of knowledge, the rank in the second-last exam captures prior knowledge with respect to the last exam and the change in rank captures her newly acquired knowledge (her progress) relative to her classmates. If, additionally, students exerted the same amount of effort for the last exam[35] this progress would reveal information about the effectiveness of a student's effort relative to her classmates. We thus expect students receiving positive change feedback to have higher confidence in the effectiveness of their effort, which raises both extrinsic motivation and intrinsic motivation. Students who receive negative change feedback will have lower confidence in the effectiveness of their effort, which decreases extrinsic motivation and lowers intrinsic motivation.

[35]The average student in our sample reported that they felt equally well prepared for the second-last and last exam (indicating average values of 2.9 for the second-last and 3.0 for the last test on a scale from 1 to 4, which are not significantly different from each other (two-sided t-test)).