Economics of Education Review 94 (2023) 102379

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

Do timing and reference frame of feedback influence high-stakes educational outcomes?[☆]

Mira Fischer [a,b,*], Valentin Wagner [c]

[a] WZB Berlin Social Science Center, Reichpietschufer 50, 10785 Berlin, Germany
[b] IZA Institute of Labor Economics, Schaumburg-Lippe-Strasse 5-9, 53113 Bonn, Germany
[c] Institute for Innovation and Technology (iit), Steinplatz 1, 10623 Berlin, Germany

ARTICLE INFO

JEL classification:
I21
D83
C93

Keywords:
Timing of feedback
Level feedback
Change feedback
High-stakes outcomes
Education

ABSTRACT

In a field experiment with fifth and sixth graders in secondary schools, students received feedback about their rank *level* or the *change* in their rank in math either a *few days* or *immediately* before the school year's final math exam. Early feedback overall enhances exam performance while late feedback worsens it. We find this pattern to hold for level feedback in general and for change feedback when negative. Heterogeneity analyses suggest the effects are mostly driven by boys, who also negatively update ability beliefs in response to feedback.

1. Introduction

Education is now more important for one's life expectancy, health, and income than ever before (Elo, 2009; Jasilionis & Shkolnikov, 2016; Michaels et al., 2014) but in spite of enormous public investments in education, about one in five students in OECD countries does not reach the baseline level of skills in mathematics, reading, and science needed to participate in society and the labor market (OECD, 2019). An important question is therefore what role students' own investments play for their educational outcomes and what motivates them to learn. Young children are particularly focused on the short term (Sutter et al., 2019) and are hardly motivated to exert effort on activities that have large long-term benefits but may not be intrinsically very enjoyable. While a number of studies have investigated the effectiveness of non-monetary and monetary incentives for younger students and have found mixed results (e.g., Bettinger, 2012; Fryer, 2011; Gneezy et al., 2019; Jalava et al., 2015; Kremer et al., 2009; Levitt et al., 2016),[1] a question that has rarely been asked is whether children respond to performance feedback that allows them to learn about their abilities but that is not itself tied to any rewards (one exemption is Beuchert et al., 2020, studying feedback in third grade). We help to fill a gap in the literature by investigating the effects of feedback to young students on their educational performance in the presence of high-stakes incentives. In particular, we ask the following questions: Is feedback about the *level* or feedback about the *change* in performance more effective? Is feedback more effective when given a *few days before* or when given *immediately before* an exam?

Studies on the "growth mindset" (Paunesku et al., 2015) and grit (Alan et al., 2019) suggest that feedback making performance *changes* over time salient may motivate students to invest effort into education. Additionally, educational policy advice from the OECD recommends against the use of marks that are based on relative performance *levels* as those depend strongly on factors that are outside a student's control, such as innate ability or the quality of a student's peers, and may discourage learning (OECD, 2012, p. 54–55). Still, students around the world mostly seem to receive the latter.[2] Furthermore, the timing of feedback possibly plays a role in its effectiveness.

[☆] Financial support by Deutsche Forschungsgemeinschaft (DFG) through CRC TRR 190 is gratefully acknowledged. We would like to thank Paul Heidhues, Matthias Heinz, Henning Hermes, Steffen Huck, Ingo Isphording, Dorothea Kübler, Johannes Münster, Gerhard Riener, Hannah Schildberg-Hörisch, Daniel Schunk, and Dirk Sliwka 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.

[*] Corresponding author at: WZB Berlin Social Science Center, Reichpietschufer 50, 10785 Berlin, Germany.
E-mail addresses: mira.fischer@wzb.eu (M. Fischer), wagner@iit-berlin.de (V. Wagner).

[1] See also Schildberg-Hörisch and Wagner (2020).

[2] "Within the context of PISA 2009, in the majority of countries and economies, students who attend higher-achieving schools receive lower marks when compared to students with similar performance and learning habits who attend poorer-performing schools" (OECD, 2012, p. 54).

https://doi.org/10.1016/j.econedurev.2023.102379
Received 6 October 2021; Received in revised form 11 February 2023; Accepted 15 February 2023
Available online 27 February 2023

Evidence shows that providing students with performance incentives immediately before an otherwise unincentivized test may significantly improve test performance (Gneezy et al., 2019), which the authors interpret as the effect of effort on the test itself. However, effort spent on deliberate practice may also strongly contribute to performance in education, and similarly at work and in sports competitions (Macnamara et al., 2014). It remains an open question whether giving feedback immediately before a high-stakes exam has a similar effect to the one found for incentives by Gneezy et al. (2019), and whether this feedback is more effective than feedback given a few days before, to which students may still adjust their preparation effort. Investigating the relative effectiveness of a widely used type of feedback (about performance levels) and of a rarely used but promising type of feedback (about performance changes) provided either just before an exam to influence effort during the exam or early enough to influence exam preparation allows us to shed light on two dimensions that are likely important for the success of feedback in education.

We use a field experiment with fifth and sixth graders to measure the effect of performance feedback given by teachers to their students on the latter's performance in a high-stakes exam.[3] Feedback was unanticipated, private, written, and given only once. The type of feedback students received was randomized within classes. 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). Across classes and within schools, we randomly varied when students received feedback: (a) a few *days* before, or (b) *immediately* before the final math exam of the semester. A survey administered immediately after the reception of feedback allows us to detect adjustments of students' ability beliefs in response to feedback.

Our results show that the timing of feedback is very important. Feedback given a few days prior to the exam significantly enhanced performance while feedback on the examination day significantly worsened performance. These effects are driven by level feedback and negative change feedback. Level feedback significantly improved performance by about 0.2 standard deviations (sd) when given early but significantly worsened performance by about 0.1 sd when given late. Change feedback – when negative – exhibits the same pattern and we find significant positive effects of 0.5 sd when given early and significant negative effects of 0.3 sd when given late. Heterogeneity analyses show that the effects are mostly driven by boys. We do not find that feedback significantly affected girls' performance. Exploring mechanisms, we find that boys tended to update beliefs negatively while girls tended to update beliefs positively in response to feedback.

Overall, our results suggest that level feedback and negative change feedback motivate young students to invest more effort into exam preparation and that this extra effort invested within a few days is highly effective in improving exam performance. Our results also caution that the same information provided immediately before an exam, when preparation is no longer possible, affects test-taking behavior negatively. Evidence from belief elicitation suggests that the reason why boys responded to feedback was because it was negatively surprising to them. This may help to explain the timing dimension: Negatively perceived information possibly motivates students to improve their exam preparation but may worsen test-taking behavior when given immediately before an exam. To the best of our knowledge, no previous studies compared the effectiveness of level and change feedback or investigated the effects of feedback timing on high-stakes outcomes. We are also not aware of any studies that investigated the effectiveness of any feedback on high-stakes outcomes for children as young as age 11–12.

Our findings are particularly relevant for educators because feedback that compares students in terms of their performance levels dominates educational practice but has been criticized for being ineffective or harmful for motivation. Results from a randomized field experiment allow us to formulate a differentiated appraisal of this claim: Level feedback may, indeed, be harmful for performance when timed badly, however, it may improve performance when timed well. Likewise, feedback about performance changes, when negative, may have positive or negative effects on performance depending on when it is given. We find some evidence that change feedback raises students' effort-effectiveness belief, which would be in line with ideas about the motivating effects of a "growth mindset" and grit. In our setting, however, responses seem to be driven by negative updating of beliefs about one's level of competence.[4] More broadly, the insight that the effects of feedback may hinge on its timing is also very relevant in other settings – such as work and professional sports – in which well-timed effort is crucial for performance.

The paper is organized as follows. The next section gives an overview of the related literature. Section 3 describes our experimental procedure. Section 4 describes the data, randomization and balance checks, and Section 5 presents the results. Section 6 concludes.

2. Related literature

While economists have studied the effects of feedback on workers and participants in laboratory experiments for some time, in recent years there has also been increasing interest in the effects of feedback on academic performance in both schools and universities (see Damgaard & Nielsen, 2018, and Villeval, 2020, for reviews of the literature). Schools are a setting in which feedback is given frequently and enhancing educational investments early in life through feedback has potentially large beneficial effects (Cunha & Heckman, 2007). However, there is only a small number of studies on feedback in schools. Azmat and Iriberri (2010) study the effect of relative performance feedback among high school students in Spain in a natural 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 relative performance feedback had positive effects and increased students' grades by 5%. However, the effect disappeared as soon as the information was removed. Andrabi et al. (2017) investigate the effects of report card dissemination with school- and child-level test scores in a field experiment in Pakistani villages and find that test scores in treatment villages increased by 0.11 sd, private school fees declined by 17% and overall enrollment increased by 4.5%. Beuchert et al. (2020) exploit the institutional design of a nationwide testing system in Danish public schools using a regression discontinuity approach. There, schools have to report to parents their child's results from the first standardized test in grade three math when children are nine to ten years old. Parents receive a letter informing them that their child's test result is "Considerably below average", "Below average", "Average", "Above average", or "Considerably above average". The authors compare the outcomes of students scoring just below a threshold, and thus receiving a more negative test feedback to

[3] We study the final math exam of the semester. Students' grade in math is a combination of the three exams during the semester and a grade for oral performance. Students in our setting need an average grade of 4 (on a scale from 1, highest, to 6, lowest) in all subjects to be promoted to the next grade and to avoid demotion to a lower academic track in Germany's rigorous early academic tracking system. Additionally, math is a core subject and if they receive a grade worse than 4 in it, they have to compensate for it by achieving at least a grade 3 in German or English.

[4] See also Fischer and Sliwka (2018), who investigate causal effects of changes in the belief about one's level of competence and the belief about the effectiveness of one's learning effort on learning behavior and test outcomes in a lab experiment.

the outcomes of students who score just above the threshold and find that students who fell just below the threshold experience a significant 0.06 sd increase in math achievement in subsequent years relative to students who scored just above the threshold. Hermes et al. (2021) study performance transparency in a mathematics e-learning application in a primary school setting in Germany. 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. In a large-scale natural experiment in which some cohorts of Greek high school students received information about their relative performance in some subjects (and not others) within their schools and across the nation, Goulas and Megalokonomou (2021) find that relative performance feedback increased high-achieving students' final-year performance by 0.15 sd whereas it led to a decrease of low-achieving students' performance by 0.3 sd.

A somewhat larger number of studies has investigated the effectiveness of feedback in higher education settings and has found mixed effects. 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. 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.[5] However, the effect vanishes one semester later. Azmat et al. (2019) provide students at a large Spanish university 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. Similarly, Dobrescu et al. (2021) investigate real-time continuous relative performance feedback which was given to students in a semester-long online assignment in a Principles of Economics course at a selective Australian university. The authors find positive feedback effects on assignments and course grades if feedback appeared when a student's rank changed. They do not find any effects for the same information being displayed continuously, only if a student's rank improved or only if a student's rank worsened. Using a sample of Japanese university students taking an introductory economics course, Kajitani et al. (2020) study the effects of providing students with information about their rank in the midterm exam on their final exam performance. They find that revealing rank information in addition to score information everyone receives improves the average scores in the final exam. They find the effect to be driven by low-performing students. Brade et al. (2022) give first-year university students in Germany (normatively framed) relative performance feedback on their accumulated course credits and find an increase in performance when the feedback is positive. However, the effect vanishes one semester later. Czibor et al. (2020) investigate the relative effectiveness of grading on the curve and an absolute grading scheme in a high-stakes testing environment among students at a Dutch university. The authors hypothesize that grading on the curve induces male students to increase their performance when compared to an absolute grading system. They find weak support for their hypothesis showing an increase in performance for the more (intrinsically) motivated male students—female students were unaffected by the grading system.

While Czibor et al. (2020) compare relative to absolute feedback, there is a literature focusing on the effectiveness of absolute performance feedback only. De Paola and Scoppa (2011) show that the frequency of absolute performance information provision matters. Students at an Italian university who received results from a mid-term exam that covered half of the course material achieve higher grades and are more likely to pass the course than students writing one exam at the end of the semester covering the full material of the course. Similarly, Pennebaker et al. (2013) study university students in Texas and find that providing them with immediate and personalized feedback by quizzes at the beginning of each lecture within a semester increases their performance compared to students taking four class-long exams over the semester. In a study investigating the effects of private, absolute feedback, Bandiera et al. (2015) exploit data of a natural experiment at a leading UK university. Feedback on exam performance improved future performance mostly for more able students and for students who initially had less information about the academic environment. Bobba and Frisancho (2022) provided disadvantaged students in Mexico with feedback on their performance on a mock version of an admission test before they had to apply for secondary schools. The authors find no effects on grades but that absolute feedback information substantially reduced the gap between perceived and actual performance.

To summarize, the literature has focused on feedback that provides students with explicit information about the level of their performance relative to others, and has found mixed or inconclusive effects overall and for different groups of students that remain largely unexplained. We are not aware of any studies investigating the relative effectiveness of level and change feedback or the role of timing for feedback effectiveness in education or other high-stakes contexts.[6]

3. Experimental procedure

3.1. Pretest of feedback notes

Before implementing the study, we tested 11–12 year old children's ability to understand the informational content of the feedback and their interpretation of it. To this end, we conducted a pretest in six classes in four schools with a total of 151 students of the same grades as our experimental sample before implementing the experiment. These children did not participate in the experiment. Details on the pretest can be found in Appendix E.

The pretest showed that most students correctly understood the feedback notes. 86% of the students could correctly calculate the implied change in the rank, and 95% could correctly determine the rank position, and 86% of students knew the exact size of their class. Students believe that motivation would be higher when receiving change feedback than when receiving level feedback while they do not indicate that the two feedback types would affect emotions differently. Overall, the results of the pretest indicate that most students of our target age group were able to understand and interpret the information given by the two types of feedback.

[5] 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 a nationwide training program in Zambia.

[6] In an economic laboratory experiment, Eriksson et al. (2009) test the effectiveness of giving feedback halfway through the time period (discrete feedback) and continuous feedback in a real-effort task in a laboratory experiment and find that feedback overall does not improve performance. In psychological laboratory experiments, 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 & Kimball, 2010, and Lechermeier & 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 (Bechtel et al., 2015; Henley & DiGennaro Reed, 2015; Krumhus & Malott, 1980). We could not find any studies that vary the interval length to the task.

3.2. Setting and preparation

The experiment was conducted in 19 fifth and sixth grade classes in seven secondary schools in the cities of Bonn, Cologne, and Düsseldorf in the most populous German federal state North Rhine-Westphalia in May and June 2016. The intervention took place just before the final math exam of the semester, which involved high stakes for the students. Students' final grade in math is a combination of the three exams during the semester and a grade for oral performance. Students in our setting need an average grade of 4 (on a scale from 1, highest, to 6, lowest) in all subjects to be promoted to the next grade. There are three core subjects—math, German, and English. If students get a grade 5 in one core subject, they can only compensate for this grade by achieving at least a grade 3 in another core subject. The experiment was approved by the ethics committee of the University of Düsseldorf and only students who received parental consent could participate in the study. Researchers were never present in the classroom to maintain a natural class setting and the feedback was given to students by their math teacher to maximize its credibility.[7] To train teachers how to conduct the experiment, we visited the schools in the run-up to the experiment. During this meeting, the intervention was explained and teachers' questions were answered. We then sent teachers two envelopes with the material needed to run the experiment. The first envelope contained written instructions for teachers, outlining the time schedule and steps of the intervention, consent forms to be signed by parents, and templates for the results of the first and the second math exam of the semester (grade and point distributions in each exam as well as the maximum number of points possible). Teachers provided us with students' names, enabling us to print personalized feedback notes by calculating students' ranks in the last math exam as well as their change in ranks from the second-to-last to the last math exam. A second envelope was sent to teachers a few days before the third exam. It contained the personalized feedback notes, which were sheets of paper that were folded and had students' names clearly written on their outside. The envelope also contained a result template for the third exam, student questionnaires, and teacher instructions that detailed how teachers should distribute feedback notes and questionnaires. They also instructed teachers to collect the questionnaires after students filled them in and required that teachers instruct their students to crumble the feedback notes and throw them in the wastepaper basket after filling in the questionnaires.[8] Upon sending the results of the final exam and the filled-out questionnaires, teachers were asked to fill out a short survey.

3.3. Treatments

Based on a 2 × 3 design, we vary both the *timing* of feedback and the *reference frame* of feedback independently to investigate how these factors influence the effectiveness of relative performance feedback on performance in a high-stakes math exam.

The timing of feedback was randomized at the class level. Students either received feedback (i) 1–3 days (usually in the last math lesson) before the exam (Early Timing) or (ii) immediately before the exam sheets were handed out (Late Timing). This treatment design allows us to investigate whether the timing of feedback matters for exam performance. The reference frame of feedback was randomized at the student level. Within the same class, students received personalized written feedback on their (I) rank level in the last math exam (Level Feedback), on their (II) change in rank between the second-last and the last math exam (Change Feedback), or (III) a personalized note that only wished them good luck for the exam (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.[9] As students had received their grades for the second last and last exam after the teacher had graded them (i.e., approximately four and two months earlier, respectively), the feedback notes contained salient and more detailed information about the level or change in their relative performance.

3.4. Questionnaire

Students in Early Timing answered an extended questionnaire after reading the feedback note that allows us to shed light on the channels through which feedback may affect students' behavioral responses. It elicited confidence in their mathematics ability, their effort-effectiveness belief, and their state self-esteem. Confidence in mathematics 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 developed a scale that asked students how much they believed (1) their correct answers and (2) their grade could be affected by their effort, then summed up scores over both questions. Their state self-esteem was measured using the Rosenberg self-esteem scale (Rosenberg, 1965). See Online Appendix J for an English translation of the scales.[10]

4. Data, randomization, and balance

We begin this section by describing the sample. Thereafter, we proceed with presenting the randomization strategy, balance checks, and descriptive statistics.

4.1. Sample

Our data consist of pre- and post-intervention math grades and exam scores, demographic information, and psychological scales from student questionnaires.

Teachers of 19 classes signed up to participate in the study. In total, 352 students in those classes received parental consent. Of those, 346 participated in the final exam. One teacher whose class was in the Early Timing treatment reported to have allowed her students to take the feedback notes and questionnaires home, thus allowing students to show them to parents, which violated our procedural instructions. We exclude this class (16 participating students) from the main analyses but include it in robustness checks. Our sample thus contains 330 students in 18 classes. Of those 330 students, 12 did not fill in a questionnaire. Of those without a questionnaire 5 were in the early treatment and 7 were in the late treatment. 9 of the 18 classes were in the early treatment and 9 in the late treatment. 15 of those 18 classes were in schools of which at least one other class participated in the experiment. In regressions that contain class fixed effects we use the whole sample of 330 students in 18 classes, however in regressions that contain school fixed effects (as robustness checks when investigating the role of timing) we can only use data of 13 classes[11] and our sample is reduced to 259 observations.

[7] 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.

[8] See Online Appendix H for English translations of the teacher instructions.

[9] See the Online Appendix I for English translations of the exact wording and layout of the notes.

[10] For exploratory heterogeneity analysis the following character traits were also elicited: locus of control (adapted from PISA [OECD, 2014]; based on Rotter, 1966), competitiveness (adapted from PISA [OECD 2014]; based on Owens & Barnes, 1992), and perseverance (adapted from PISA [OECD 2014]). In Late Timing, due to time constraints, students could only fill in a short questionnaire after completing the exam. It contained some of the subscales of some of the scales for exploratory analysis.

[11] One school only had one participating class. Two classes of another school are in the late treatment but none is in the early treatment after we had

4.2. Randomization and balance checks

For schools that had several participating classes, classes were randomized into either the Early Timing treatment or the Late Timing treatment blocked on school level. Within classes, students were then randomized into the Control group, Change Feedback treatment or Level Feedback treatment at the individual level. Balance tests show that between student-level treatments all baseline-measures were perfectly balanced, see Tables A.1 and A.2 in Appendix A. For class-level treatments we also achieve good balance for baseline performance, gender, and being old relative to class level,[12] however the proportion of students with migration background (marginally not significant) and with siblings (significant at the 5%-level) tends to be higher in early-feedback classes. We will show regressions with and without demographic control variables to assess the effects of these small imbalances on our results. There are no significant differences in variables that vary at the class level, such as the proportion of female teachers, class size, the share of participating students, and the share of sixth graders between the class-level treatments. Tables A.3 and A.4 in Appendix A report balance tests between Early Timing and Late Timing.

4.3. Descriptive statistics

47% of the students in our sample are female, 38% of students have a non-German first and family name, which we use as a proxy for migration background,[13] 13% of students are older than expected for their grade, and 86% of students have one or more siblings. The average grade in math prior to the intervention is 2.6 on a scale from 1 to 6, where 1 is the highest and 6 is the lowest grade.[14] Class-level variables show that the overwhelming majority of teachers is female, each class has about 27 students, 68% of which participate in the experiment, and two thirds of students in our sample attend fifth grade while one third attends sixth grade. 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 F.1 and F.2 in the Online Appendix show the distribution of feedback pooled over class-level treatments.

5. Results

This section begins with a presentation of the effects of feedback on performance, including a heterogeneity analysis, and proceeds with an exploration of potential mechanisms. We conclude the section by considering possible spillover effects.

Table 1
Descriptive statistics of provided feedback.

		Obs.	Mean	Std. Dev	Min.	Max.
Change Feedback	Early Timing	53	.89	7.93	−21	21
	Late Timing	56	.79	8.30	−19	19
Level Feedback	Early Timing	58	14.64	8.30	1	30
	Late Timing	58	13.50	8.22	1	30
Control	Early Timing	50	–	–	–	–
	Late Timing	55	–	–	–	–

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

5.1. Effects of feedback on performance

In this section we investigate the effects of the timing and the type of feedback on students' performance in the final exam. First, we study the effect of feedback *timing* on performance pooling over feedback types. Thereafter, we show results for each feedback type when feedback is given a few days prior to the final exam or on the day of the final exam. Next, we investigate whether heterogeneous reactions to feedback exist by gender, feedback valence, and feedback magnitude. Finally, we explore potential mechanisms by which feedback could affect outcomes, and discuss possible spillover effects.

Effects of feedback timing (pooled over type)

To investigate the effect of feedback on performance and to shed light on the role of timing, we estimate linear regression models separately for each feedback timing, pooling over feedback types and controlling for class fixed effects. Furthermore, as we expect the grade in the final exam and the response to feedback to both strongly depend on one's prior grade in math, we control for it using fixed effects that correspond to the major categories in the German grading scale (prior grade = {1,2,3,4,5,6}). Furthermore, to control for slight imbalances in some of the demographic variables between early and late classes, we include dummy variables for gender, migration background, being old relative to one's class level, and having siblings in the model. The reported standard errors are clustered at the class level and corrected using bias-reduced linearization (Angrist & Pischke, 2008; Bell & McCaffrey, 2002; Cameron et al., 2008; Cameron & Miller, 2015) to allow for cluster-robust inference with a small number of clusters. For each feedback timing we estimate the following model:

$$\begin{aligned} Grade\,Final\,Exam_i &= \beta_0 + \beta_1 Feedback_i + \beta_2 Prior\,Grade_i \\ &\quad + \beta_3 Demographics_i + \beta_4 Class_j + \varepsilon_{ij} \end{aligned} \tag{1}$$

In Table 2, we can see that providing students with (any) feedback about their past performance a few days prior to the exam has a positive effect on their performance, while providing the same kind of feedback immediately before the exam decreases their performance. Looking at columns 3 and 6, which contain the full models for Early Timing and Late Timing classes, respectively, we can see that giving feedback a few days before the exam has large positive and significant effect on students' grades of 0.224 standard deviations (sd) ($p = 0.031$), while giving feedback immediately before the exam decreases performance by 0.126 sd (p = 0.075). As an alternative specification we estimate models with school fixed effects that control for heterogeneity at the school level (reducing our sample size by five classes to 259 observations) and introduce an interaction term of feedback with timing. Table B.1 in Appendix B shows that this gives us very similar estimates. This table also shows that students in the control group whose peers get feedback a few days prior to the exam do not perform differently in the final exam than students in the control group whose peers get feedback on the examination day, as the coefficient of *Late* is not significant. We further discuss this result in Section 5.3 where we consider possible spillover effects on the control group.

to exclude the class in which the experimental instructions were violated. Furthermore, two classes of another school are both in the early treatment but none is in the late treatment because one teacher dropped out after class-level randomization but did not respond to the material we sent them and did not provide us with any information on their class.

[12] This affects students who had to repeat a class or whose transition from kindergarten to primary school was deferred because of developmental delays or at parents' request.

[13] 41.5% of secondary school students in North Rhine-Westphalia, the German federal state in which the experiment was conducted, have a migration background. See: https://www.schulministerium.nrw.de/docs/bp/Ministerium/Service/Schulstatistik/Amtliche-Schuldaten/StatTelegramm2016.pdf.

[14] 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 2
Effects of feedback timing on performance (Pooled over Type).

	Early			Late		
	(1)	(2)	(3)	(4)	(5)	(6)
Feedback	0.120	0.224**	0.224**	−0.224**	−0.114	−0.126*
	(0.091)	(0.096)	(0.102)	(0.107)	(0.069)	(0.070)
Class FE	Yes	Yes	Yes	Yes	Yes	Yes
Prior Performance	No	Yes	Yes	No	Yes	Yes
Dem. Controls	No	No	Yes	No	No	Yes
Adj. R^2	0.094	0.405	0.427	0.093	0.339	0.342
N	161	161	152	169	169	157

Note: This table shows the effects of feedback on performance using a linear regression model. The dependent variable is the grade in the final exam (inverted, such that larger grades are better), standardized to mean zero and standard deviation one. *Feedback* identifies students who received any feedback (either Change Feedback or Level Feedback). Columns 1–3 show results for classes in which the treatment groups received feedback a few days prior to the exam (Early), and columns 4–6 show results for classes in which the treatment groups received feedback immediately before the exam (Late). All models contain class fixed effects and a constant. Models in columns 2 and 4 additionally control for prior performance using fixed effects that correspond to the major categories in the German grading scale (six categories, five dummies). Models in columns 3 and 6 additionally control for demographics: gender, migration background, being old relative to one's class level, having siblings. Standard errors are reported in parentheses, clustered at class level, and corrected using bias-reduced linearization. The number of clusters is 9 in all models. * p<0.10, ** p<0.05, *** p<0.01.

Table 3
Effects of feedback timing and type on performance.

	Early			Late		
	(1)	(2)	(3)	(4)	(5)	(6)
Change Feedback	0.106	0.207	0.202	−0.190	−0.082	−0.081
	(0.110)	(0.141)	(0.132)	(0.133)	(0.078)	(0.096)
Level Feedback	0.132	0.240***	0.244**	−0.258**	−0.145	−0.167**
	(0.101)	(0.074)	(0.096)	(0.116)	(0.099)	(0.074)
Class FE	Yes	Yes	Yes	Yes	Yes	Yes
Prior Performance	No	Yes	Yes	No	Yes	Yes
Dem. Controls	No	No	Yes	No	No	Yes
Adj. R^2	0.088	0.401	0.423	0.088	0.335	0.339
N	161	161	152	169	169	157

Note: This table shows the effects of feedback on performance using a linear regression model. The dependent variable is the grade in the final exam (inverted, such that larger grades are better), standardized to mean zero and standard deviation one. *Change Feedback* identifies students who received Change Feedback and *Level Feedback* identifies students who received Level Feedback. Columns 1–3 show results for classes in which the treatment groups received feedback a few days prior to the exam (Early), and Columns 4–6 show results for classes in which the treatment groups received feedback immediately before the exam (Late). All models contain class fixed effects and a constant. Models in columns 2 and 4 additionally control for prior performance using fixed effects that correspond to the major categories in the German grading scale (six categories, five dummies). Models in columns 3 and 6 additionally control for demographics: gender, migration background, being old relative to one's class level, having siblings. Standard errors are reported in parentheses, clustered at class level, and corrected using bias-reduced linearization. The number of clusters is 9 in all models. * p<0.10, ** p<0.05, *** p<0.01.

Effects of feedback type (by timing)

We now investigate the role of the reference frame of feedback by introducing separate dummies for the Change Feedback treatment and the Level Feedback treatment. Otherwise, we estimate the same model as above. For each feedback timing we estimate the following model:

$$\begin{aligned} Grade\,Final\,Exam_i = {} & \beta_0 + \beta_1 Change\,Feedback_i + \beta_2 Level\,Feedback_i \\ & + \beta_3 Prior\,Grade_i + \beta_4 Demographics_i + \beta_5 Class_j + \varepsilon_{ij} \end{aligned} \tag{2}$$

Table 3 presents treatment effect estimates of the two feedback types by the timing of feedback. Column 3 shows that level feedback has a positive effect of 0.244 sd (p = 0.012) when given early and column 6 shows that it has a negative effect of 0.167 sd (p = 0.026) when given late. While the effects of change feedback show the same sign reversal with respect to timing (Early: β = 0.202, p = 0.127; Late: β = −0.081, p = 0.403) neither of the effects for change feedback are significant.

We find the results for both feedback types and timings to be robust to the inclusion of the non-compliant class in the early treatment and to controlling for prior performance using linear control variables for prior grades (see Table G.1 and Table G.2 in Online Appendix G).

Heterogeneity analysis

Feedback valence and magnitude. In this section we briefly summarize and discuss exploratory heterogeneity analyses by feedback valence and magnitude. We investigate whether students who received level feedback about a rank in the upper half respond differently than students whose feedback contained a rank in the lower half with respect to their class. We also test whether students respond to incremental changes of level feedback. Similarly, we investigate whether students who received feedback about negative changes in performance respond differently to change feedback than students whose change feedback reported an improvement. Furthermore, we test whether students respond to incremental changes of feedback about negative and positive change. Note that all these analyses are supposed to provide suggestive evidence only as cell size is reduced to about 25 observations when we introduce interaction terms.

Table 4 suggests that the positive effects of level feedback given early are driven by students who receive feedback about a rank in the lower half. These students' performance increases by 0.358 sd (p = 0.054), while the coefficient for students who receive feedback about a rank in the upper half is close to zero and insignificant (column 1). Furthermore, we find that students respond to incremental changes in level feedback given early. As the rank a student receives feedback about increases (i.e. gets worse) by one, performance increases by 0.019 sd (p = 0.060, column 2). We do not detect significant heterogeneous effects for level feedback when given late (columns 3 and 4).

Furthermore, Table C.1 in Appendix C suggests that negative change feedback when given early has a significant positive effect of 0.461 sd on performance (p = 0.003, column 3) and that negative change feedback when given late has a marginally significant negative effect of 0.263 sd on performance (p = 0.080, column 7). We also find that in the late treatment students react more strongly to negative change feedback the more negative it is. As the negative change a student learns about increases by one, performance decreases by 0.181 sd (p = 0.000, column 8). We do not find any significant effects for positive change feedback given early or late (columns 1–2 and 5–6).

Overall, these findings suggest that students mostly react to negative feedback, i.e. feedback about a negative change or a low rank, and that negative feedback tends to increase performance when given early but to lower performance when given late. We also find some evidence that students react more strongly to more extreme feedback.

Gender. Next, we briefly report the results of a heterogeneity analysis by gender, which due to small cell sizes should be interpreted as suggestive evidence only. It has widely been shown that male subjects are more confident of their abilities than female subjects (Barber & Odean, 2001; Niederle & Vesterlund, 2007). Thus, when receiving feedback about their past performance boys might have to adjust their beliefs about their abilities more negatively than girls, potentially leading to heterogeneity in responses to feedback.

The results in Table 5 confirm that boys respond differently to feedback than girls. Boys who receive any type of feedback a few days prior to the exam experience a significant increase in their performance. Change feedback given early to boys significantly improves their performance by 0.376 sd (p = 0.018), similarly, level feedback given early to them significantly improves their performance by 0.299 sd (p = 0.000, column 1). When boys receive change feedback late, their performance tends to decrease (change feedback: β = −0.244, p = 0.293; level feedback: β = −0.173,p = 0.318) but we do not find

Table 4
Effects of feedback type and timing on performance (Interaction with half and rank).

	Early		Late	
	(1)	(2)	(3)	(4)
Change Feedback	0.473*	−0.334	−0.188	0.082
	(0.241)	(0.299)	(0.142)	(0.289)
Level Feedback	0.358*	−0.005	−0.157	−0.231
	(0.185)	(0.166)	(0.231)	(0.166)
Change × Better Half	−0.623**		0.131	
	(0.302)		(0.276)	
Level × Better Half	−0.296		−0.087	
	(0.206)		(0.310)	
Better Half	0.557***		0.398	
	(0.154)		(0.271)	
Change × Prior Rank		0.037*		−0.016
		(0.022)		(0.019)
Level × Prior Rank		0.019*		0.003
		(0.010)		(0.015)
Prior Rank		−0.042***		−0.041**
		(0.008)		(0.019)
Class FE	Yes	Yes	Yes	Yes
Prior Performance	Yes	Yes	Yes	Yes
Dem. Controls	Yes	Yes	Yes	Yes
Adj. R^2	0.435	0.444	0.352	0.377
N	152	152	157	157

Note: This table shows the effects of feedback type on performance by the timing of feedback using a linear regression model. Columns 1 and 2 present results for students receiving feedback a few days prior to the exam and columns 3 and 4 present results for students receiving feedback immediately before the exam. The dependent variable is the grade in the final exam (inverted, such that larger grades are better), standardized to mean zero and standard deviation one. All models contain class fixed effects, a constant, and the full set of control variables: mean prior grade rounded to the nearest integer over two previous math exams (5 dummies), gender, migration background, being old relative to one's class level, having siblings. Models in columns 1 and 3 additionally contain a dummy *Better Half,* identifying students who performed in the upper half (i.e. in terms of the absolute grade, where smaller grades are better) relative to their classmates in the last math exam prior to the intervention, and interactions of this indicator with the treatment dummies (*Change* × *Better Half* and *Level* × *Better Half*). Models in columns 2 and) additionally contain a variable *Prior Rank,* capturing a student's rank in the last math exam prior to the intervention, and interactions of this indicator with the treatment dummies (*Change* × *Prior Rank* and *Level* × *Prior Rank*). Standard errors are reported in parentheses, clustered at class level, and corrected using bias-reduced linearization. The number of clusters is 9 in all models. * p<0.10, ** p<0.05, *** p<0.01.

Table 5
Effects of feedback type and timing on performance (By gender).

	Early	Late
	(1)	(2)
Change Feedback	0.376**	−0.244
	(0.157)	(0.231)
Level Feedback	0.463***	−0.173
	(0.083)	(0.173)
Change Feedback × Female	−0.401*	0.322
	(0.240)	(0.445)
Level Feedback × Female	−0.461**	0.019
	(0.213)	(0.413)
Class FE	Yes	Yes
Prior Performance	Yes	Yes
Dem. Controls	Yes	Yes
Adj. R^2	0.425	0.334
N	152	157

Note: This table shows the effects of feedback type on performance by the timing of feedback using a linear regression model. Column 1 presents results for students receiving feedback a few days prior to the exam and column 2 presents results for students receiving feedback immediately before the exam. The dependent variable is the grade in the final exam (inverted, such that larger grades are better), standardized to mean zero and standard deviation one. All models contain class fixed effects, a constant, and the full set of control variables: mean prior grade rounded to the nearest integer over two previous math exams (5 dummies), gender, migration background, being old relative to one's class level, having siblings. Additionally, all models contain interactions of *Female* with the treatment dummies (*Change Feedback* × *Female* and *Level Feedback* × *Female*). Standard errors are reported in parentheses, clustered at class level and corrected using bias-reduced linearization. The number of clusters is 9 in all models. * p<0.10, ** p<0.05, *** p<0.01.

Table 6
Summary of significant effects for confidence, self-esteem, and effort-effectiveness belief.

	Sign. effect on:	Boys	Girls
Change Feedback	Confidence		
Level Feedback		–	+
Change Feedback	Self-Esteem		+
Level Feedback		–	
Change Feedback	Effort-Effect. Belief		
Level Feedback		–	

Note: This table summarizes the results of our exploration of mechanisms. "–" indicates that a significant negative effect was found, while "+" indicates that a significant positive effect was found. Empty cells indicate that no significant effect was found.

the effects to be significant (column 2). In contrast to this, for girls the estimated coefficients for both feedback types and both timings are very close to zero and insignificant.[15]

5.2. Exploration of potential mechanisms

In this section, we summarize and discuss our findings with respect to feedback effects on students' beliefs. Table D.1 suggests that neither feedback type affected students' math confidence overall (column 1) and we do not find significant effects by feedback valence (columns 2 and 3). A heterogeneity analysis by gender reveals that level feedback significantly decreased boys' confidence in their math ability by 0.191 sd (p = 0.008) and increased girls' confidence in their math ability by 0.384 sd (p = 0.047, column 4). Table D.2 suggests that neither feedback type affected students' self-esteem overall (column 1) and we do not find significant effects by feedback valence (columns 2 and 3). A heterogeneity analysis by gender reveals that both change and level feedback tended to decrease boys' self-esteem, by 0.495 (p = 0.116) sd and 0.398 (p = 0.001) sd, respectively. Change feedback increased girls' self-esteem by 0.497 sd (p = 0.060, column 4). Table D.3 suggests that change feedback had an overall positive effect of 0.304 (p = 0.84) on students' effort-effectiveness belief (column 1) and we do not find significant effects by feedback valence (columns 2 and 3) or by gender (column 4). However, level feedback decreases boys' effort-effectiveness belief by 0.400 sd (p = 0.046, model 4). As the heterogeneity analysis by gender is based on small cell sizes and we are looking at multiple, non-incentivized belief outcomes, these results, too, should be interpreted as suggestive evidence only.

Taking into account that these results should be treated with caution, we tentatively interpret the above results as suggesting that boys overall tended to update beliefs negatively in response to feedback while girls tended to update beliefs positively, suggesting that the former were negatively surprised while the latter were positively surprised by the information they were provided with. These findings are summarized in Table 6 below and are in line with the heterogeneous effects of feedback on performance summarized in the end of Section 5.1. The findings suggest that early feedback motivated exam preparation and

[15] When we interact gender with feedback valence, we observe, consistently, that boys tend to respond more positively to negative change feedback in Early Timing and more negatively to negative change feedback in Late Timing, respectively, than girls. However, cell sizes become very small and we do not detect any significant effects.

late feedback demotivated effort during the exam of boys because they were negatively surprised by it. Some theoretical considerations about how the effect of feedback may depend on prior ability beliefs and feedback timing can be found in Appendix D.2.

5.3. Possible spillover effects

Our main 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. In late timing classes students could not find out anything about the feedback other students had received as students were already seated separately to write the exam and received sheets formatted in the same way. Thus, as students could not talk about the content of their sheets and control group students did not know that other students received feedback, spillover effects were not possible in Late Timing. However, in early timing classes negative and positive spillover effects of our intervention on students who did not receive any feedback are possible as students in the treatment group could have talked to students in the control group between the intervention and writing the exam (even if the notes could not be shown to other students because they had to be destroyed after filling in the questionnaire). Possibly, students in the control group who found out that their classmates received feedback 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 Change and Level Feedback in Early Timing classes to be overestimated. Alternatively, students in the control group 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.

As can be seen in Table B.1, the coefficient of "Late", which measures by how much the performance of control group students of Late Timing differs from the performance of control group students in Early Timing, is slightly negative and insignificant. We therefore infer that the spillover effects of our intervention on the control group were, if anything, positive and that the positive coefficients of level and change feedback in classes who received feedback early are therefore not inflated by negative spillovers on the control group.

6. Conclusion

We investigate the effects of feedback on young students' performance in a high-stakes mathematics exam. We vary two dimensions that likely matter for its effects: timing and reference frame. As both preparation and test-taking behavior may contribute to performance, students receive feedback at one of two points in time when it may be expected to have the largest effects on the respective behaviors: a few days before or immediately before the exam. Teachers either gave students feedback that captures their performance level relative to their classmates or feedback that informs students about the change in their performance relative to their classmates.

Our results show that feedback overall improves performance when given early but decreases performance when given late. We find this pattern to hold for level feedback in general and for change feedback that is negative. Early feedback particularly benefits the performance of boys, who on average seem to be disappointed by the feedback, suggesting that they were overconfident in the absence of feedback. Overall, our results suggest that performance feedback, in particular when it is negative or disappointing, motivates better exam preparation and that this extra effort invested within a few days is highly effective in improving math performance. However, our results also caution that the same information provided immediately before an exam, when preparation is no longer possible, influences test-taking behavior negatively.

Our study contributes to the literature in the economics of education that investigates interventions to raise student motivation and the literature in organizational economics that studies the effects of feedback on performance. To the best of our knowledge, no previous studies investigated the effects of feedback timing on high-stakes outcomes, educational or other, and no previous studies compared the relative effectiveness of level and change feedback. Our results are particularly relevant for educators because performance feedback is routinely given in the classroom but the widely used practice of comparing students' performance levels has been criticized for being ineffective or even harmful (see OECD, 2012). We find this claim to hold only for feedback given immediately before an exam but not for feedback given earlier and we detect no clear benefits for change feedback over level feedback in this respect. In fact, as shown by our study, the timing of feedback may be of paramount importance, an insight that is also relevant in other contexts, such as sports and job performance, in which the ability to motivate people to invest effort into preparation and skill acquisition is crucial.

Table A.1
Balance check student-level treatments – Early Timing.

	(1) Control	(2) Change	(3) Level	(4) Overall	(5) (1) vs. (2), p-value	(6) (1) vs. (3), p-value	(7) (2) vs. (3), p-value
Prior Performance	2.626 (0.158)	2.689 (0.160)	2.754 (0.146)	2.693 (0.089)	0.781	0.552	0.762
Female	0.420 (0.071)	0.415 (0.068)	0.517 (0.066)	0.453 (0.039)	0.960	0.317	0.286
Migrant	0.480 (0.071)	0.434 (0.069)	0.362 (0.064)	0.422 (0.039)	0.643	0.219	0.444
Older	0.120 (0.046)	0.120 (0.046)	0.170 (0.052)	0.137 (0.028)	1.000	0.479	0.479
Siblings	0.900 (0.043)	0.918 (0.040)	0.909 (0.039)	0.909 (0.023)	0.754	0.876	0.868
N	50	53	58	161			
Proportion	0.311	0.329	0.360	1			

Note: This table reports group means of key characteristics of students for the student-level treatments (Control, Change, Level) of Early Timing classes in columns 1–3. *Prior Performance* is the average grade over the two previous, *Female* indicates that a student is female, *Migrant* indicates whether students have a migration background, *Older* indicates that a student is old relative to their class level, and *Siblings* indicates that a student has siblings. 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 A.2
Balance check student-level treatments – Late Timing.

	(1) Control	(2) Change	(3) Level	(4) Overall	(5) (1) vs. (2), p-value	(6) (1) vs. (3), p-value	(7) (2) vs. (3), p-value
Prior Performance	2.455 (0.135)	2.633 (0.138)	2.671 (0.124)	2.588 (0.076)	0.360	0.241	0.839
Female	0.418 (0.067)	0.536 (0.067)	0.483 (0.066)	0.479 (0.039)	0.219	0.495	0.576
Migrant	0.382 (0.066)	0.286 (0.061)	0.345 (0.063)	0.337 (0.036)	0.287	0.686	0.502
Older	0.125 (0.048)	0.151 (0.050)	0.107 (0.042)	0.127 (0.027)	0.710	0.779	0.499
Siblings	0.800 (0.057)	0.811 (0.054)	0.839 (0.050)	0.818 (0.031)	0.886	0.603	0.704
N	55	56	58	169			
Proportion	0.325	0.331	0.343	1			

Note: This table reports group means of key characteristics of students for the student-level treatments (Control, Change, Level) of Late Timing classes in columns 1–3. *Prior Performance* is the average grade over the two previous,*Female* indicates that a student is female, *Migrant* indicates whether students have a migration background, *Older* indicates that a student is old relative to their class level, and *Siblings* indicates that a student has siblings. 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 A.3
Balance check class-level treatments (student characteristics).

	(1) Late timing	(2) Early timing	(3) Overall	(4) (1) vs. (2), p-value
Prior Performance	2.588 (0.076)	2.693 (0.089)	2.639 (0.058)	0.370
Female	0.479 (0.039)	0.453 (0.039)	0.467 (0.028)	0.639
Migrant	0.337 (0.036)	0.422 (0.039)	0.379 (0.027)	0.112
Older	0.127 (0.027)	0.137 (0.028)	0.132 (0.019)	0.798
Siblings	0.818 (0.031)	0.909 (0.023)	0.863 (0.019)	0.019
N	169	161	330	
Proportion	0.512	0.488	1	

Note: This table reports group means of key characteristics of students for the Late Timing (column 1) and Early Timing (column 2) treatments. Column 3 presents means for the pooled sample. *Prior Performance* is the average grade over the two previous, *Female* indicates that a student is female, *Migrant* indicates whether students have a migration background, *Older* indicates that a student is old relative to their class level, and *Siblings* indicates that a student has siblings. 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.

CRediT authorship contribution statement

Mira Fischer: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Validation, Writing – original draft, Writing – review & editing. **Valentin Wagner:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Validation, Writing – original draft, Writing – review & editing.

Data availability

The data that has been used is confidential.

Appendix A. Balance checks

See Tables A.1–A.4.

Appendix B. Early versus late feedback: Alternative specification

See Table B.1.

Table A.4
Balance check class-level treatments (class characteristics).

	(1) Late timing	(2) Early timing	(3) Overall	(4) (1) vs. (2), p-value
Teacher Female	0.778 (0.147)	0.778 (0.147)	0.778 (0.101)	1.000
Class Size	26.889 (1.338)	27.111 (1.195)	27.000 (0.871)	0.903
Share Participated	0.704 (0.072)	0.661 (0.057)	0.683 (0.045)	0.648
Sixth Grade	0.333 (0.167)	0.333 (0.167)	0.333 (0.114)	1.000
N	9	9	18	
Proportion	0.500	0.500	1	

Note: This table reports group means of key characteristics of the class for the Late Timing (column 1) and Early Timing (column 2) treatments. Column 3 presents means for the pooled sample. *Teacher Female* indicates that the class's math teacher is female, *Class Size* is the total number of students in a class, *Share Participated* is the share of students in a class that participated in the experiment, and *Sixth Grade* indicates indicates that a class is in sixth grade. 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.

Appendix C. Heterogeneity analysis

See Table C.1.

Appendix D. Potential mechanisms

D.1. Effects of feedback on beliefs

See Tables D.1–D.3.

D.2. Theoretical considerations

In the following, we outline a simple framework –speculative but inspired by existing evidence –that may explain why we found positive effects for Early but negative effects for Late feedback on exam performance. Villeval (2020) argues that feedback may have positive effects, through e.g., more accurate beliefs about the marginal returns to effort, or negative effects, through e.g., stress, on performance. There is ample evidence that more accurate beliefs allow people to make more optimal

Table B.1
Effects of feedback timing on performance (Pooled over Type) – Alternative specification.

	(1)	(2)	(3)
Feedback	0.145	0.220**	0.226**
	(0.105)	(0.104)	(0.103)
Feedback * Late	−0.342**	−0.316**	−0.386***
	(0.163)	(0.138)	(0.148)
Late	−0.126	−0.167	−0.147
	(0.173)	(0.110)	(0.134)
School FE	Yes	Yes	Yes
Prior Performance	No	Yes	Yes
Dem. Controls	No	No	Yes
Adj. R^2	0.069	0.369	0.373
N	259	259	242

Note: This table shows the effects of feedback timing on performance using a linear regression model. The dependent variable is the grade in the final exam (inverted, such that larger grades are better), standardized to mean zero and standard deviation one. *Feedback* indicates whether students received any feedback (either Change Feedback or Level Feedback) and *Late* identifies classes in which the treatment groups received feedback immediately before the exam. All models contain school fixed effects and a constant. The model in columns 2 additionally controls for prior performance using fixed effects that correspond to the major categories in the German grading scale (six categories, five dummies). The model in column 3 additionally controls for demographics: gender, migration background, being old relative to one's class level, having siblings. Standard errors are reported in parentheses, clustered at class level, and corrected using bias-reduced linearization. The number of clusters is 13 in all models. * p<0.10, ** p<0.05, *** p<0.01.

effort decisions (Santos-Pinto & de la Rosa, 2020). Additionally, stress has been shown to influence decision making under uncertainty (Duque et al., 2022), to affect competition behavior (Buser et al., 2017), to increase subjective discounting, and to decrease effort (Delaney et al., 2014).

Behavior driven by "emotional" stress responses is more likely to be observed in the short run, while behavior driven by "rational" belief updating is more likely to be observed in the longer run, possibly because reference points shift and the information is processed differently (Gneezy & List, 2006; Loewenstein, 2005; Loewenstein & Schkade, 1999). The following exposition combines the above ideas and shows how the assumption that short-run responses to feedback are dominated by emotions while long-run responses are dominated by rational belief updating may explain the opposite effects of early and late feedback.

The economic literature on education assumes that students' exam performance is influenced by school, parental, and student inputs (e.g., Behrman et al., 2015; Hanushek, 2003). As our intervention targets student inputs, the following exposition focuses on them.

We assume that a student's utility positively depends on his exam grade (G), which is a function of his ability (a), his effort during learning and during the test (e), and his concentration during the test (t):

$u = G(a, e, t)$, with $\frac{\partial G}{\partial a} > 0$, $\frac{\partial G}{\partial e} > 0$, and $\frac{\partial G}{\partial t} > 0$.

Students have beliefs about their ability, which we call "confidence in ability" ($\hat{a}$). Both learning effort and concentration during the exam depend on confidence in ability: $e(\hat{a})$ and $t(\hat{a})$. Confidence shocks are processed rationally if a behavioral response is only required in the longer term such that they only affect incentives to exert effort. Confidence shocks are processed emotionally if a behavioral response in required immediately and may affect the effectiveness at a given task by affecting short-term concentration. In Early Timing students can take their time to process and respond to the feedback. Furthermore, we assume that learning effort decreases as confidence in ability increases: $\frac{\partial e}{\partial \hat{a}} < 0$. The better a student believes his math ability already is, the smaller is the necessity to invest effort into learning for the exam in order to get his desired grade. (Similarly, effort during the exam may decrease but the time delay between feedback in Early Timing and taking the exam make it likely that there are no or negligent effects of early feedback on effort during the exam and that learning effort is the main channel.) In Late Timing students have to respond immediately. Furthermore, we assume concentration during the exam depends positively on confidence in ability: $\frac{\partial t}{\partial \hat{a}} > 0$. The better a student believes his math ability is, the less anxious he feels during the exam and the better he is able to concentrate.

If a student is *overconfident* he overestimates his true performance and Level Feedback about his past performance is *negatively* surprising. Negative Change Feedback is likely negatively surprising independent

Table C.1
Effects of feedback type and timing on performance by change direction (Interaction with magnitude of change).

	Early				Late			
	Pos. change		Neg. change		Pos. change		Neg. change	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Change Feedback	−0.041	−0.278	0.461***	0.218	0.003	0.234	−0.263*	0.883**
	(0.415)	(0.437)	(0.148)	(0.347)	(0.210)	(0.426)	(0.148)	(0.349)
Level Feedback	0.176	0.020	0.341*	0.158	−0.094	−0.290	−0.273	−0.103
	(0.285)	(0.407)	(0.194)	(0.426)	(0.174)	(0.317)	(0.182)	(0.288)
Change × Change in Rank (abs.)		0.045		0.038		−0.019		−0.187***
		(0.041)		(0.049)		(0.048)		(0.051)
Level × Change in Rank (abs.)		0.028		0.026		0.037		−0.005
		(0.037)		(0.043)		(0.039)		(0.042)
Change in Rank (abs.)		−0.040		0.006		−0.033		0.098**
		(0.031)		(0.033)		(0.044)		(0.038)
Class FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Prior Performance	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Dem. Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adj. R^2	0.324	0.296	0.548	0.548	0.282	0.276	0.368	0.463
N	82	82	70	70	75	75	82	82

Note: This table shows the effects of feedback type on performance by the timing of feedback using a linear regression model. Columns 1–4 present results for students receiving feedback a few days prior to the exam and columns 5–8 present results for students receiving feedback immediately before the exam. Columns 1, 2, 5, and 6 presents results for students whose performance improved between the second-last and the last math exam and columns 3, 4, 7, 8 present results for students whose performance worsened between the second-last and the last math exam. The dependent variable is the grade in the final exam (inverted, such that larger grades are better), standardized to mean zero and standard deviation one. All models contain class fixed effects, a constant, and the full set of control variables: mean prior grade rounded to the nearest integer over two previous math exams (5 dummies), gender, migration background, being old relative to one's class level, having siblings. Models in columns 2, 4, 6, and 8 additionally contain a variable *Change in Rank (abs.),* capturing a student's change in their rank from the second last to the last math exam, and interactions of this variable with the treatment dummies (*Change Feedback* × *Change in Rank (abs.)* and *Level Feedback* × *Change in Rank (abs.)*). Standard errors are reported in parentheses, clustered at class level and corrected using bias-reduced linearization. The number of clusters is 9 in all models. * p<0.10, ** p<0.05, *** p<0.01.

Table D.1
Effects of feedback type on math confidence.

	(1)	(2)	(3)	(4)
Change Feedback	−0.009 (0.201)	0.330 (0.211)	−0.209 (0.239)	−0.058 (0.228)
Level Feedback	0.099 (0.084)	0.083 (0.194)	−0.008 (0.150)	−0.191*** (0.069)
Change Feedback × Pos. Change		−0.618*** (0.127)		
Level Feedback × Pos. Change		0.016 (0.327)		
Pos. Change		0.259 (0.285)		
Change Feedback × Better Half			0.416 (0.380)	
Level Feedback × Better Half			0.245 (0.304)	
Better Half			−0.272 (0.332)	
Change Feedback × Female				0.120 (0.431)
Level Feedback × Female				0.593*** (0.212)
Class FE	Yes	Yes	Yes	Yes
Prior Performance	Yes	Yes	Yes	Yes
Dem. Controls	Yes	Yes	Yes	Yes
Adj. R^2	0.211	0.216	0.200	0.217
N	148	148	148	148

Note: This table shows the effects of feedback type on students' math confidence using a linear regression model. The dependent variable is a scale measuring confidence in mathematics ability, standardized to mean zero and standard deviation one. All models contain class fixed effects, a constant, and the full set of control variables: mean prior grade rounded to the nearest integer over two previous math exams (5 dummies), gender, migration background, being old relative to one's class level, having siblings. The model in column 2 additionally contains a dummy *Pos. Change* identifying students whose rank improved between the second last and the last math exam, and interactions of this indicator with the treatment dummies (*Change Feedback × Pos. Change* and *Level Feedback × Pos. Change*). The model in column 3 additionally contains a dummy *Better Half,* identifying students who performed in the upper half relative to their classmates in the last math exam prior to the intervention, and interactions of this indicator with the treatment dummies (*Change Feedback × Better Half* and *Level Feedback × Better Half*). The model in column 4 additionally contains interactions of *Female* with the treatment dummies (*Change Feedback × Female* and *Level Feedback × Female*). Standard errors are reported in parentheses, clustered at class level and corrected using bias-reduced linearization. The number of clusters is 9 in all models. * p<0.10, ** p<0.05, *** p<0.01.

of prior confidence due to its clear direction ("worsened"). Negatively surprising feedback in turn decreases confidence in ability.

If a student is *underconfident* he underestimates his true performance and Level Feedback about his past performance is *positively* surprising. Positive Change Feedback is likely positively surprising independent of prior confidence due to its clear direction ("improved"). Positively surprising feedback in turn increases confidence in ability.

Giving negatively (positively) surprising feedback on the examination day (Late Timing) decreases (increases) short-term concentration and hence decreases (increases) a student's performance. On the contrary, giving negatively (positively) surprising feedback a few days before the exam (Early Timing), increases (decreases) learning effort because it increases (decreases) the incentive to do so and hence increases (decreases) a student's performance.

The expected effects of feedback timing based on the above reasoning can be found in Table D.4.

Thus, whether Early Timing and Late Timing feedback may be expected to have positive or negative effects on performance depends on whether students were negatively or positively surprised by the feedback.

Table D.2
Effects of feedback type on state self-esteem.

	(1)	(2)	(3)	(4)
Change Feedback	−0.124 (0.117)	0.118 (0.190)	−0.206 (0.347)	−0.612* (0.321)
Level Feedback	−0.266* (0.149)	−0.111 (0.243)	−0.155 (0.145)	−0.402*** (0.103)
Change Feedback × Pos. Change		−0.475 (0.301)		
Level Feedback × Pos. Change		−0.304 (0.378)		
Pos. Change		0.430 (0.290)		
Change Feedback × Better Half			0.182 (0.539)	
Level Feedback × Better Half			−0.230 (0.369)	
Better Half			−0.122 (0.269)	
Change Feedback × Female				1.143** (0.513)
Level Feedback × Female				0.347 (0.322)
Class FE	Yes	Yes	Yes	Yes
Prior Performance	Yes	Yes	Yes	Yes
Dem. Controls	Yes	Yes	Yes	Yes
Adj. R^2	0.126	0.122	0.113	0.171
N	144	144	144	144

Note: This table shows the effects of feedback type on students' state self-esteem using a linear regression model. The dependent variable is a scale measuring state self-esteem, standardized to mean zero and standard deviation one. All models contain class fixed effects, a constant, and the full set of control variables: mean prior grade rounded to the nearest integer over two previous math exams (5 dummies), gender, migration background, being old relative to one's class level, having siblings. The model in column 2 additionally contains a dummy *Pos. Change* identifying students whose rank improved between the second last and the last math exam, and interactions of this indicator with the treatment dummies (*Change Feedback × Pos. Change* and *Level Feedback × Pos. Change*). The model in column 3 additionally contains a dummy *Better Half,* identifying students who performed in the upper half relative to their classmates in the last math exam prior to the intervention, and interactions of this indicator with the treatment dummies (*Change Feedback × Better Half* and *Level Feedback × Better Half*). The model in column 4 additionally contains interactions of *Female* with the treatment dummies (*Change Feedback × Female* and *Level Feedback × Female*). Standard errors are reported in parentheses, clustered at class level and corrected using bias-reduced linearization. The number of clusters is 9 in all models. * p<0.10, ** p<0.05, *** p<0.01.

Table D.3
Effects of feedback type on effort effectiveness belief.

	(1)	(2)	(3)	(4)
Change Feedback	0.330** (0.163)	0.459* (0.263)	0.205 (0.308)	0.158 (0.176)
Level Feedback	−0.046 (0.187)	−0.150 (0.239)	−0.266 (0.346)	−0.400** (0.198)
Change Feedback × Pos. Change		−0.244 (0.516)		
Level Feedback × Pos. Change		0.180 (0.374)		
Pos. Change		−0.010 (0.277)		
Change Feedback × Better Half			0.223 (0.500)	
Level Feedback × Better Half			0.450 (0.654)	
Better Half			−0.103 (0.441)	

(*continued on next page*)

Table D.3 (*continued*).

	(1)	(2)	(3)	(4)
Change Feedback × Female				0.386 (0.513)
Level Feedback × Female				0.724*** (0.259)
Class FE	Yes	Yes	Yes	Yes
Prior Performance	Yes	Yes	Yes	Yes
Dem. Controls	Yes	Yes	Yes	Yes
Adj. R^2	0.093	0.089	0.092	0.110
N	150	151	151	151

Note: This table shows the effects of feedback type on students' effort effectiveness belief using a linear regression model. The dependent variable is a scale measuring a student's effort effectiveness belief, standardized to mean zero and standard deviation one. All models contain class fixed effects, a constant, and the full set of control variables: mean prior grade rounded to the nearest integer over two previous math exams (5 dummies), gender, migration background, being old relative to one's class level, having siblings. The model in column 2 additionally contains a dummy *Pos. Change* identifying students whose rank improved between the second last and the last math exam, and interactions of this indicator with the treatment dummies (*Change Feedback* × *Pos. Change* and *Level Feedback* × *Pos. Change*). The model in column 3 additionally contains a dummy *Better Half,* identifying students who performed in the upper half relative to their classmates in the last math exam prior to the intervention, and interactions of this indicator with the treatment dummies (*Change Feedback* × *Better Half* and *Level Feedback* × *Better Half*). The model in column 4 additionally contains interactions of *Female* with the treatment dummies (*Change Feedback* × *Female* and *Level Feedback* × *Female*). Standard errors are reported in parentheses, clustered at class level and corrected using bias-reduced linearization. The number of clusters is 9 in all models. * p<0.10, ** p<0.05, *** p<0.01.

Table D.4
Summary: Effects of feedback timing assuming different short- and long-run responses.

	Late Timing	Early Timing
If negatively surprising	Confidence in ability ↓ Concentration ↓ Performance ↓	Confidence in ability ↓ Learning Effort ↑ Performance ↑
If positively surprising	Confidence in ability ↑ Concentration ↑ Performance ↑	Confidence in ability ↑ Learning Effort ↓ Performance ↓

Appendix E. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.econedurev.2023.102379.

References

Alan, S., Boneva, T., & Ertac, S. (2019). Ever failed, try again, succeed better: Results from a randomized educational intervention on grit. *Quarterly Journal of Economics*, *134*(3), 1121–1162.

Andrabi, T., Das, J., & Khwaja, A. I. (2017). Report cards: The impact of providing school and child test scores on educational markets. *American Economic Review*, *107*(6), 1535–1563.

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

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

Azmat, G., Bagues, M., Cabrales, A., & Iriberri, N. (2019). What you don't know... Can't hurt you? A field experiment on relative performance feedback in higher education. *Management Science*, *65*(8), 3714–3736.

Azmat, G., & 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.

Bandiera, O., Larcinese, V., & 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.

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

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

Behrman, J. R., Parker, S. W., Todd, P. E., & Wolpin, K. I. (2015). Aligning learning incentives of students and teachers: Results from a social experiment in Mexican High Schools. *Journal of Political Economy*, *123*(2), 325–364.

Bell, R., & McCaffrey, D. (2002). Bias reduction in standard errors for linear and generalized linear models with multi-stage samples. *Survey Methodology*, *28*, 169–179.

Bettinger, E. (2012). Paying to learn: The effect of financial incentives on elementary school test scores. *The Review of Economics and Statistics*, *94*(3), 686–698.

Beuchert, L., Eriksen, T. L. M., & Krægpøth, M. V. (2020). The impact of standardized test feedback in math: Exploiting a natural experiment in 3rd grade. *Economics of Education Review*, *77*, Article 102017.

Bobba, M., & Frisancho, V. (2022). Self-perceptions about academic achievement: Evidence from Mexico City. *Journal of Econometrics*, *231*(1), 58–73.

Brade, R., Himmler, O., & Jäckle, R. (2022). Relative performance feedback and the effects of being above average — Field experiment and replication. *Economics of Education Review*, *89*, Article 102268.

Buser, T., Dreber, A., & Mollerstrom, J. (2017). The impact of stress on tournament entry. *Experimental Economics*, *20*(2), 506–530.

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

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

Cunha, F., & Heckman, J. (2007). The technology of skill formation. *American Economic Review*, *97*(2), 31–47.

Czibor, E., Onderstal, S., Sloof, R., & Van Praag, M. (2020). Does relative grading help male students? Evidence from a field experiment in the classroom. *Economics of Education Review*, *75*, Article 101953.

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

De Paola, M., & Scoppa, V. (2011). Frequency of examinations and student achievement in a randomized experiment. *Economics of Education Review*, *30*(6), 1416–1429.

Delaney, L., Fink, G., & Harmon, C. P. (2014). *Effects of stress on economic decision-making: Evidence from laboratory experiments*: *Discussion paper 8060*, Institute for the Study of Labour (IZA).

Dobrescu, L. I., Faravelli, M., Megalokonomou, R., & Motta, A. (2021). Relative performance feedback in education: Evidence from a randomised controlled trial. *The Economic Journal*, *131*(640), 3145–3181.

Duque, A., Cano-López, I., & Puig-Pérez, S. (2022). Effects of psychological stress and cortisol on decision making and modulating factors: A systematic review. *European Journal of Neuroscience*, *56*(2), 3889–3920.

Elo, I. T. (2009). Social class differentials in health and mortality: Patterns and explanations in comparative perspective. *Annual Review of Sociology*, *35*(1), 553–572.

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

Fischer, M., & 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.

Fryer, R. G. (2011). Financial incentives and student achievement: Evidence from randomized trials. *Quarterly Journal of Economics*, *126*(4), 1755–1798.

Gneezy, U., & List, J. (2006). Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica*, *74*(5), 1365–1384.

Gneezy, U., List, J. A., Livingston, J. A., Qin, X., Sadoff, S., & Xu, Y. (2019). Measuring success in education: The role of effort on the test itself. *American Economic Review: Insights*, *1*(3), 291–308.

Goulas, S., & Megalokonomou, R. (2021). Knowing who you actually are: The effect of feedback on short- and longer-term outcomes. *Journal of Economic Behaviour and Organization*, *183*, 589–615.

Hanushek, E. (2003). The failure of input-based schooling policies. *The Economic Journal*, *113*(485), F64 – F98.

Henley, A. J., & DiGennaro Reed, F. 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., & Schunk, D. (2021). Motivating low-achievers — Relative performance feedback in primary schools. *Journal of Economic Behaviour and Organization*, *187*, 45–59.

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

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

Jasilionis, D., & Shkolnikov, V. M. (2016). Longevity and education: A demographic perspective. *Gerontology*, *62*, 253–262.

Kajitani, S., Morimoto, K., & Suzuki, S. (2020). Information feedback in relative grading: Evidence from a field experiment. *PLoS One*, *15*(4), Article e0231548.

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

Krumhus, K. M., & 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.

Lechermeier, J., & 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.

Levitt, S., List, J., Neckermann, S., & 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. (2005). Hot-cold empathy gaps and medical decision making. *Health Psychology*, *24*(4), 49–56.

Loewenstein, G., & Schkade, D. (1999). Wouldn't it be nice? Predicting future feelings. In *Well-being: The foundations of hedonic psychology* (pp. 85–105).

Macnamara, B. N., Hambrick, D. Z., & Oswald, F. L. (2014). Deliberate practice and performance in music, games, sports, education, and professions: A meta-analysis. *Psychological Science*, *25*(8), 1608–1618.

Michaels, G., Natraj, A., & Van Reenen, J. (2014). Has ICT polarized skill demand? Evidence from eleven countries over twenty-five years. *The Review of Economics and Statistics*, *96*(1), 60–77.

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

OECD (2012). *Grade expectations: How marks and education policies shape students' ambitions*. PISA, OECD Publishing.

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

OECD (2019). *PISA 2018 results (volume I): What students know and can do*. Paris: OECD Publishing.

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

Paunesku, D., Walton, G., Romero, C., Smith, E., Yeager, D., & Dweck, C. (2015). Mind-set interventions are a scalable treatment for academic underachievement. *Psychological Science*, *26*(6), 784–793.

Pennebaker, J. W., Gosling, S. D., & Ferrell, J. D. (2013). Daily online testing in large classes: Boosting college performance while reducing achievement gaps. *PLoS One*, *8*(11), 1–6.

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

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

Santos-Pinto, L., & de la Rosa, L. E. (2020). Overconfidence in labor markets. In K. F. Zimmermann (Ed.), *Handbook of labor, human resources and population economics* (pp. 1–42). Cham: Springer International Publishing.

Schildberg-Hörisch, H., & Wagner, V. (2020). Monetary and non-monetary incentives for educational attainment: Design and effectiveness. In S. Bradley, & G. Green (Eds.), *The economics of education: A comprehensive overview (second edition)* (pp. 249–268). Elsevier.

Smith, T. A., & 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.

Sutter, M., Zoller, C., & Glätzle-Rützler, D. (2019). Economic behavior of children and adolescents - A first survey of experimental economics results. *European Economic Review*, *111*, 98–121.

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

Villeval, M. C. (2020). Performance feedback and peer effects. In K. F. Zimmermann (Ed.), *Handbook of labor, human resources and population economics*. Springer.