Hermes, Henning; Huschens, Martin; Rothlauf, Franz; Schunk, Daniel

Conference Paper

Motivating Low-Achievers – Relative Performance Feedback in Primary Schools

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2020: Gender Economics

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Motivating Low-Achievers—Relative Performance Feedback in Primary Schools

Henning Hermes*[1], Martin Huschens[2], Franz Rothlauf[2], and Daniel Schunk[3]

[1]NHH Bergen, FAIR / Department of Economics
[2]University of Mainz, Information Systems and Business Administration
[3]University of Mainz, Public and Behavioral Economics

November 2019

Abstract

Relative performance feedback (RPF) has often been shown to improve effort and performance in the workplace and educational settings. Yet, many studies also document substantial negative effects of RPF, in particular for low-achievers. We study a novel type of RPF designed to overcome these negative effects of RPF on low-achievers by scoring individual performance *improvements*. With a sample of 400 children, we conduct a class-wise randomized-controlled trial using an e-learning software in regular teaching lessons in primary schools. We demonstrate that this type of RPF significantly increases motivation, effort, and performance in math for low-achieving children, without hurting high-achieving children. Among low-achievers, those receiving more points and moving up in the ranking improved strongest on motivation and math performance. In addition, we document substantial gender differences in response to this type of RPF: improvements in motivation and learning are much stronger for girls. We argue that using this new type of RPF could potentially reduce inequalities, especially in educational settings.

Keywords: relative performance feedback, rankings, randomized-controlled trial, education, gender differences

*Corresponding author: henning.hermes@nhh.no; NHH Bergen, Helleveien 30, 5045 Bergen, Norway

1 Introduction

Information about peer behavior and outcomes is crucial for social comparison processes (Festinger, 1954), social reference point formation[1], and for the perception of social norms.[2] Consequently, relative performance feedback has been shown to have a substantial effect on individual perceptions, choices, and behavior, thus often being a strong and sustainable motivator for human beings. In work-related contexts, evidence shows that, even in the absence of pecuniary incentives, relative performance feedback (e.g., based on rankings) can lead to increased motivation, effort, and work performance (Gill, Kissova, Lee, & Prowse, 2018; Hannan, McPhee, Newman, & Tafkov, 2013; Blanes i Vidal & Nossol, 2011). In other words, people are highly interested in their social ranking ("rank incentives", Barankay, 2012). Thus, relative performance feedback is considered a low-cost instrument for increasing motivation, effort, and performance in firms (Blanes i Vidal & Nossol, 2011).[3]

In a similar vein, the provision of feedback, including relative performance feedback, is often used in educational settings (Hattie & Timperley, 2007). Ample research has analyzed the effects of relative performance feedback in education on effort provision, test performance, and learning outcomes (Bursztyn & Jensen, 2015; Ashraf, Bandiera, & Lee, 2014; Azmat, Bagues, Cabrales, & Iriberri, 2019; Azmat & Iriberri, 2010; Tran & Zeckhauser, 2012; Megalokonomou & Goulas, 2018; Fischer & Wagner, 2018; Brade, Himmler, & Jäckle, 2018). Tran and Zeckhauser (2012) provided evidence that students who received rank information showed performance increases in standardized tests. Similarly, two studies reported that performance feedback may lead to increased test performance by (i) supporting students to form well-informed self-appraisals and providing additional information on how effort translates into outcomes (Bandiera, Larcinese, & Rasul, 2015) and (ii) activating an energizing competitive drive within students who gain utility (disutility) from being ahead (behind) of others (Azmat & Iriberri, 2010).

On the other hand, relative performance feedback might also entail substantial costs in the form of negative effects on low-achievers. This could potentially exaggerate exist-

[1] Behavior and outcomes of peers are a reasonable source of individual reference point formation (cf. Haenni (2019) for a recent empirical study of this phenomenon). In particular, observing others (or: receiving information about others) may influence, e.g., how individuals perceive fairness (Fehr & Schmidt, 1999; Bolton & Ockenfels, 2000) and judge their subjective well-being (Veblen, 1899; Card, Mas, Moretti, & Saez, 2012).

[2] Individuals derive social norms from observing others and, in consequence, alter behavior and choices accordingly (see, for example, Cialdini, Reno, & Kallgren, 1990; Coleman, 1990)

[3] A related strand of research deals with rank-order tournaments and the design of optimal labor contracts in which compensation is based on relative rank rather than absolute levels of output (see Lazear & Rosen, 1981, for an early contribution). Connelly, Tihanyi, Crook, and Gangloff (2014) provided a meta-analytic review of tournament theory in management research. Empirical evidence on the effects of relative performance feedback that is tied to tournament pay schemes is inconclusive: results range from detrimental effects on performance and effort (Delfgaauw, Dur, Sol, & Verbeke, 2013; Hannan, Krishnan, & Newman, 2008) to effort-enhancing effects (Eriksson, Poulsen, & Villeval, 2009; Lazear & Rosen, 1981; Prendergast, 1999).

ing inequalities—an effect which is very likely to be undesirable in educational settings. Indeed, when looking at the effect of relative performance feedback on performance and learning outcomes, some studies found mixed or even detrimental effects, especially for low-achievers: Bursztyn and Jensen (2015) found that introducing a performance leader board into computer-based high school courses resulted in a severe performance decline across all ability groups due to a strong desire to avoid the leader board. Megalokonomou and Goulas (2018) showed that disclosing information on students' relative high school performance led to a performance improvement for high-achieving students (by about 0.15 SD), while the performance of low-achievers dropped (by 0.30 SD). Similarly, also Ashraf et al. (2014) reported negative effects of rankings and social comparison on effort in a health worker training program due to "self-handicapping" processes of low-ability individuals. In addition to that, Azmat et al. (2019) found evidence of detrimental effects of the provision of relative performance feedback on students' learning outcomes: negative effects were mainly driven by those subjects who initially underestimated their performance level and subsequently showed lower effort.[4]

Understanding these differences in reaction to relative performance feedback along the ability distribution is an important question in itself. In light of the frequent use of relative performance feedback in educational settings[5] and its potential negative effects on low-achievers (and, as a result, on inequality), understanding the effects of relative performance feedback and how to improve relative performance feedback in educational settings is of even higher relevance. Hence, we contribute to the literature in several ways: First, we propose a novel type of relative performance feedback, namely scoring individual performance *improvements* in contrast to absolute performance; we designed our method to overcome previously evidenced negative effects of relative performance feedback on low-achievers. Second, we analyze the effects of this new type of relative performance feedback by conducting a randomized-controlled field study in primary schools (in contrast to many natural experiments in this literature)—thereby, we achieve a maximum of control while exogenously varying the type of feedback received in an externally valid setting.

[4] Moreover, several studies showed negative effects of relative performance feedback on outcomes in the area of risk-taking (e.g., Linde & Sonnemans, 2012; Dijk, Holmen, & Kirchler, 2014; Kirchler, Weitzel, & Lindner, 2018) and social behavior (e.g., Charness, Masclet, & Villeval, 2014; Kuziemko, Buell, Reich, & Norton, 2014)—this was not the focus of the present study but we report some (null) results on these areas in the SOM, see Section A.2.

[5] The introduction of e-learning systems into schools that apply game-based feedback systems relying on relative performance feedback (point systems, rankings, high-score lists, etc.) accelerates the prevalence of relative performance feedback in educational settings even more (gamification, e.g., Sailer, Hense, Mayr, & Mandl, 2017; Deterding, Dixon, Khaled, & Nacke, 2011), and it has been noted that it is in particular certain features of digital learning environments such as instantaneous or visual feedback provision that are conducive to the effectiveness of these tools for improving learning outcomes (see, e.g., der Kleij, Feskens, & Eggen, 2015; Dobrescu, Faravelli, Megalokonomou, & Motta, 2019).

Third, we also contribute to filling a gap in this literature by providing *field evidence* on the effects of *repetitive* and *continuous* relative performance feedback over a period of several weeks in an externally valid classroom setting.Therefore, we can account for the dynamic effects of relative performance feedback provided over time. Fourth and finally, we examine a broad range of important outcomes in the educational context, measured in a highly standardized computer-based way, such as motivation, effort, and learning outcomes (and also some "softer" outcomes such as (self-reported) perceived stress, self-efficacy, and liking of competition). Hence, we are able to provide a comprehensive picture of the effects of relative performance feedback and to better understand channels and mechanisms of potential treatment effects.

To this end, we introduced a mathematics e-learning software package into primary schools, applying a class-wise randomized-controlled trial (RCT) design. Class-wise randomization generates lower statistical power than within-class randomization; it is, however, the only feasible way of implementing such a study without creating fundamental spill-over problems and having children learn about their treatment condition (which would create a whole range of severe methodological problems). Treatment and control group used the same e-learning software, with an identical user interface, the same content, and the same frequency and saliency of feedback. The e-learning software in the treatment group only differed with regard to the provision of *relative* performance feedback in form of a ranking, while children in the control group received solely private, *individual* performance feedback (see Figure 1). Feedback in both groups, treatment *and* control, was based on a point system that relied on scoring *individual performance improvements* rather than giving points for *absolute performance*. The concept of this type of feedback was developed to mirror pedagogical guidelines in primary school that encourage teachers to evaluate individual improvements in contrast to absolute performance, in order to put them into a growth mindset (cf., e.g., Claro, Paunesku, & Dweck, 2016).

Importantly, treatment and control group use exactly the same point system—thus, differences in outcomes can only be attributed to the fact that children in the treatment group continuously received *relative* performance feedback about their peers in the form of a ranking, while the control group only learned about *individual* performance. Note that this experimental design does not allow us to infer insights on the effects of providing feedback on performance *improvements* compared to (classical) absolute performance feedback. In contrast, with this study we want to take a first step in learning about the effect of *relative* performance feedback using feedback on performance *improvements* in order to overcome

Figure 1: Mathematics E-Learning Software in Treatment and Control Condition

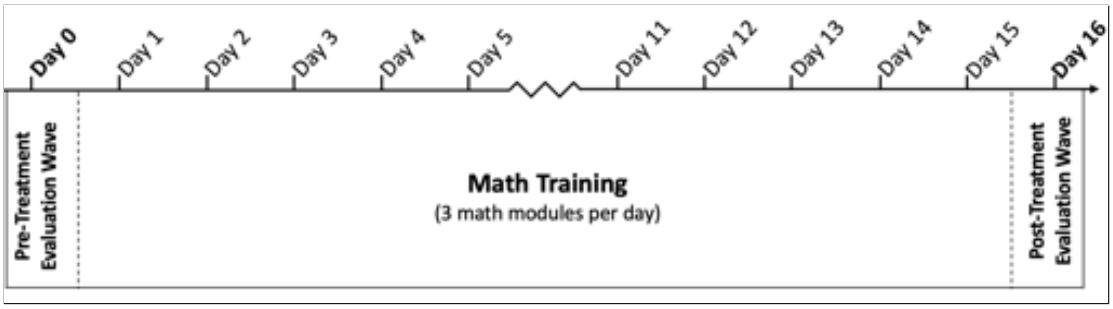
Notes: The left screenshot shows the treatment group condition with *relative performance feedback*. This feedback condition consisted of a leader board indicating the rank, the individually designed avatar, the first name, and the number of points gained by all children in the class. The right screenshot shows the control group condition with *individual performance feedback*. This feedback condition only consisted of the child's individually designed avatar, the name, and the individual number of points.

the observed negative effects of *classical* relative performance feedback on low-achievers.[6] Note that under the hypothesis that feedback on performance *improvements* has motivating effects for low-achieving children, any treatment effect of *relative* performance feedback we observe in our setting would come on top of this (hypothetical) basic motivation effect of scoring performance improvements (because children in the control group also received feedback on performance improvements).

About 400 third-graders in 20 classes at seven primary schools in Germany used this mathematics e-learning software over the course of five weeks for 15 lessons (see Figure 2). We randomized on class-level and within schools. Each lesson consisted of a compulsory part in which several math modules were first practiced (without the possibility of earning points) and then evaluated for points (in a test mode called "*Sprint*"). Points were based on individual performance improvements compared to previous *Sprints*. After that, children had the opportunity to voluntarily practice the math tasks (without the possibility of earning points). We measured the outcomes of our study directly before (pre-treatment evaluation wave) and after the five-week treatment period (post-treatment evaluation wave) using highly standardized, objective test measures as well as teacher and child questionnaires.

[6] An alternative experimental design would compare classical relative performance feedback (i.e., scoring absolute performance) with our new type of relative performance feedback. While we acknowledge that this design could yield interesting insights, we deliberately decided not to use it, for two main reasons: On the theoretical side, we did not want our new type of feedback points to be confounded with the *relative* performance feedback. In other words, comparing classical relative performance feedback with our new relative performance feedback would yield differences in both, individual feedback (i.e., many or few points received) as well as the relative ranking of children in class (and the dynamic development). Our experimental design can account for this by shutting down the channel of differences in points received: children in treatment and control group receive exactly the same individual feedback, the only thing we add for the treatment group is relative ranking information. The second reason is on the practical side: The majority of teachers (at least in the primary schools in Germany we recruited for our experiment) expressed a very negative attitude towards "classical" relative performance feedback in our very first discussions about the study. Thus, to avoid the problem that differences in teachers' attitudes drive treatment effects (because teachers could not be blind to treatment in a setting like this), we decided to compare our new type of feedback using individual feedback vs. relative performance feedback. We also considered that a third treatment cell was not feasible due to limited statistical power.

Figure 2: Timeline of the Field Study



Notes: Each day represents one school lesson of 50 minutes. During the math training, children had to do math tasks in 11 different math modules (M1–M11); please refer to Figure S10 for an overview of the math modules). Each module was repeated four times on four different days. A more detailed version of this timeline can be found in the SOM, Figure S1.

Additionally, we also collected data on children's behavior while they used the mathematics e-learning software.

Our main results are as follows: Despite our limited statistical power, we are able to detect significant improvements for low-achieving children in the treatment group compared to the control group. Our new type of relative performance feedback strongly boosted motivation and effort for low-achievers. Further, this increase in motivation and effort seems to translate into improvements in actual learning outcomes, namely math performance. Importantly, these positive effects for low-achievers are not associated with costs (i.e., lower performance) for middle- and high-achievers. Among low-achievers, those children that (i) receive higher numbers of points or that (ii) can improve their ranking over time show strongest increases, supporting the notion that the relative performance feedback (i.e., the ranking) is the key driver of our treatment effects on low-achieving children. Self-reported ratings by the children indicate an increase in perceived stress for the low-achievers in the treatment group but also show increases in self-efficacy in math and higher liking (i.e., a more positive attitude) of competition in general. Interestingly, we report strong gender differences in reactions to our new type of relative performance feedback: girls adjust their motivation and effort, and thus show strong improvements in math performance (but also report higher perceived stress), while boys report higher self-efficacy and a more positive attitude toward competition. Overall, our findings suggest that relative performance feedback about performance improvements could be a powerful tool to ameliorate inequalities, especially in educational settings.

The remainder of the paper is structured as follows: First, we report our exact experimental procedures (Section 2), followed by our results (Section 3). We discuss our findings in Section 4 and conclude in Section 5. Further details, figures, and tables can be found in the Supplementary Online Material (SOM, Sections A–C).

2 Materials and Methods

2.1 Procedures

We recruited seven primary schools with 20 third-year classes and around 400 children for participation in the study. At the beginning of the school year, third-year children worked for 15 school lessons (five weeks, three lessons per week, in regular teaching lessons) with a mathematics e-learning software package that we specifically developed for this study. All lessons were conducted by trained research assistants in a highly standardized manner, using game-like interactions embedded in the software and audio-visual, automated instructions. Every child was seated in front of an individual notebook computer, used an external mouse to interact with the software, and had a headset to listen to the audio instructions.

The study consisted of three parts: two evaluation waves (pre- and post-treatment measurement) and a training phase (see Figure S1). This design allowed us to (i) increase the precision of our estimates by including baseline scores for each respective outcome, and (ii) to better control for non-perfect randomization. Baseline measurements took place in one school lesson prior to the first training lesson (but within the same week); post-treatment measurements also lasted for one school lesson and were conducted after the last training lesson (but within the same week).

In the training phase, children could earn points for their achievements in the math tasks. In contrast, evaluation waves were incentivized with "gold coins" (note that the gold coins were only used in the evaluation waves; they did not influence the point system or the ranking used in the training phase—neither did the number of points have any influence on the number of golden coins children received). Gold coins could be used as currency to buy one toy at the end of the study, with a larger number of coins allowing the child to choose from a larger selection of toys.

2.2 Participants

When our study began, there were 404 children in the 20 participating classes. We were able to gain consent from 399 parents for study participation, resulting in a participation rate of 98.8%. On average, children were 8.61 years old (SD = 0.48); 53% of the children were male. Class size ranged from 18 to 24 children. Of our 399 children in the sample, we had to exclude 16 from our analyses because teachers and/or our research assistants indicated that their language level was not sufficient to understand or use the e-learning

software for the study.[7] Therefore, the final sample size was n = 383; however, for five children there was information missing about their grade (Math or German) and, thus, in our main results, sample size is reduced to n = 378. Table S3 in the SOM reports the complete sample characteristics of our study.

2.3 Randomization

Randomization was implemented (i) at the class level, (ii) within schools, and (iii) stratified based on the socioeconomic status (SES) of the school district. Table S4 shows that the randomization process was successful, as the experimental groups did not differ on important sociodemographic variables. Importantly, there was also no significant difference in the number of low-achieving or high-achieving children between the treatment and the control group. Comparing baseline levels of the outcome variables for treatment and control group (see Tables S5–S6), we also see that for all outcomes but one (Math Multiplication/Division) randomization was successful. Given the number of outcomes measured, it is not surprising that we found one outcome variable to be significantly different between our experimental groups (nonetheless, we control for baseline scores of the respective outcomes in all our estimations).

2.4 Treatment

We designed and developed the mathematics e-learning software specifically for the present study. It consisted of 11 math modules that repeated and practiced the curriculum of the second grade (basic arithmetic operations). Over a period of five weeks, children trained three days per week, one school lesson (50 min) per day. On each training day, they used three different math modules, with each module consisting of a prescribed training phase to practice on the task, a testing phase (*Sprint*) that was relevant to gather points, and a voluntary practice phase in which children could do additional training tasks. In total, every math module was repeated four times on four different days (*Sprints 1–4*).

Both treatment and control group used exactly the same learning software with the same user interface, functionalities, weekly schedule, and math modules. The software differed only in one single feature, which was the relative component of the performance feedback. The control group was presented only with individual performance feedback,

[7] As reported by our research assistants, most of these children were refugees and had arrived in Germany only recently. For some of them, teachers actually allowed them to use the software, whereas other teachers gave them different tasks to work on during the training lessons. However, we decided to exclude all these children from our analyses because many outcome tasks including the questionnaires required a certain level of language proficiency which was clearly not met by these children.

i.e., children were permanently shown the cumulative points that they achieved by solving math problems in the test mode (*Sprints*). In contrast, the software in the treatment group provided children—in addition to the individual performance feedback—with a permanently visible and dynamic ranking of all children in their class, showing individual points and ranks for all children. The ranking was dynamic in a sense that it was constantly updated during the school lesson in which children were using the software. In both groups, feedback was displayed prominently on the right-hand side of the screen to provide the feedback information as saliently as possible. Screenshots of the experimental conditions can be found in Figure 1.

2.5 Point System

The basis for the implementation of performance feedback in both experimental groups during the math training was a point system based on children's *performance improvements* (measured as time improvements) over the four repetitions of each math module (i.e., *Sprints 1–4*). Thus, the higher the improvement, the more points the child could earn.

To be able to measure and compare performance improvements across tasks, children, and classrooms, we had to identify an easy and continuous measure of performance in math tasks. We decided to use the time children needed to correctly solve a given set of tasks. Solving these type of simple math tasks quickly and without errors is an important learning goal in primary school and is a key prerequisite for acquiring more advanced math skills. Note that providing a wrong answer caused a waiting time; hence, children had no incentive to guess without doing the calculations first.

In a first step (*Sprint 1*), we evaluated the individual baseline performance in each math module measured as the absolute time children needed to finish a set of tasks. The *Sprint* was successfully finished if all tasks were answered correctly within the given time frame of 180 seconds. The ranking for *Sprint 1* worked as follows: the fastest child in class was ranked first and received 10 points, while the slowest child was ranked last and received one point. In between, children were given points (integer numbers) based on their relative rank in the class distribution. If ties occurred, children received the same number of points. Consequently, the baseline measurement (*Sprint 1*) ranked children according to their *absolute* performance.

The following three *Sprints* (i.e., *Sprints 2–4*) in each math module were used to elicit performance *improvements* and rank children accordingly: the child who improved most compared with his or her average previous performance was ranked first and received 10

points. Children who did not improve or stagnated with regard to the time they needed were assigned one point. Similar to *Sprint 1*, children were given between 1 and 10 points based on their relative rank in the class distribution.

Children collected points cumulatively over the period of five weeks, i.e., they received feedback about the number of points (1–10) achieved from a specific *Sprint* and these points were added to the points collected in previous rounds. The display on the right-hand side of the screen (see Figure 1) showed this total number of points (control group) or the public leader board with all total numbers of points in class in descending order (treatment group). While collecting the points cumulatively over time likely increases external validity (as it mirrors how these leader boards and ranking are usually constructed), one might be worried that dynamics in rank changes slow down over time. We analyzed this looking at the mean standard deviation of average ranks over time (and across subgroups of children); Figure S9 in the SOM reports that, while dynamics slow down after the first three days of the intervention (as one would expect with a cumulative point system), there is substantial dynamics in ranking up until the end of the intervention period. This is most likely driven by the continuous introduction of new math modules which generate new potentials for improvements for all children. Also, there is no difference in ranking dynamics across the three subgroups of children, i.e., low-achievers have (on average) the same standard deviation in average rank as middle- or high-achieving children.

2.6 Outcome Measures

We collected a broad range of outcome measures pre- and post-treatment in a highly standardized and incentivized way. To measure the treatment effects on children's *motivation*, we used a computer-based motivation task designed to capture intrinsic motivation, teacher ratings on children's motivation, self-rated motivation (as rated by the child), as well as the number of tasks solved voluntarily within the e-learning software and the time spent on these tasks (note that, in contrast to all other measures, voluntary practice tasks are measured *during* the intervention). In order to measure transfer effects on *learning outcomes* we used two conceptually different sets of math tasks, namely addition and subtraction as well as multiplication and division. The two sets of math tasks were designed to be very different from the training tasks used during the intervention but, at the same time, aimed to measure exactly the math competencies trained during the intervention (see Figures S10 and S11 in the SOM for the differences). Finally, we report results for self-rated outcomes in a child questionnaire for perceived stress, somatic problems, self-efficacy in mathematics,

and their liking of competition in general. A detailed description of data collection methods and outcomes can be found in the SOM in Section A.3.

To identify low-achievers (or high-achievers) in school, we needed information on a broad range of school-related abilities and behaviors of a child based on a long period of time. Hence, we derived our classification into low- and high-achievers based on teacher-reported grades (at baseline) as they integrate information on children's school achievements over time. We classify children with a "bad" grade prior to treatment (lower than 3 on a scale from 1 (very good) to 6 (insufficient)) as "low-achieving" children (n = 99) and children with a very good grade (i.e., 1) prior to treatment as "high-achieving" children (n = 114). The remaining children are classified as "middle-achieving" children (n = 165).

2.7 Data Analysis

We use OLS regressions to estimate treatment effects. Specifically, we regress post-treatment levels for our outcomes on a dummy variable indicating treatment status. To analyze the treatment effect across the ability distribution, we include treatment interaction terms for low- and high-achieving children as well as dummies for low- and high-achieving children. Thus, the reference category is middle-achieving children in the control group. In addition, we control for school fixed effects, gender, age, pretreatment grades in Math and German (as rated by the teacher), and baseline levels of the respective outcome[8] to increase precision. We cluster standard errors at the class level and because we have only a small number of clusters (i.e., 20 classes), we use a conservative correction method known as "biased-reduced linearization" (BRL, Bell & McCaffrey, 2002), which is more conservative than the standard cluster-robust variance estimator. To compare effects across outcomes, all outcomes were standardized to mean = 0 and SD = 1 (see a short description of outcomes in Section 2.6 and full details in the SOM (Section A.3). Data analysis was conducted using Stata 15 SE and R (version 3.2.5).

3 Results

We identify the effect of our new type of relative performance feedback by comparing children of low-, middle- and high-ability type in a classroom who received relative performance feedback, with children in a classroom who received only individual performance feedback (of the same type), within the same school.

[8] Note that there is no baseline score for voluntary practice tasks during training. We did not collect baseline scores for perceived stress as rated by the child. Hence, in the OLS regression analyses of these cases, we cannot control for baseline scores.

We begin by analyzing treatment effects on motivation and effort (see Table 1). For the motivation test task and teacher-rated motivation, we find a strong and significant heterogeneous treatment effect for low-achieving children (0.47 SD ($p = .034$) for the motivation test, 0.43 SD ($p = .008$) for teacher-rated motivation). The linear combination of the interaction effect with the treatment dummy, indicating the difference between the subgroup of low-achieving children in the treatment group and the low-achieving children in the control group, is substantial and highly significant—apparently, low-achieving children in the treatment group perform 0.34 SD ($p = .026$) better in the motivation task and are rated 0.47 SD ($p = .094$) higher on motivation by their teachers. Effects on middle- and high-achieving children are smaller and statistically insignificant. Child-rated motivation seems somewhat increased for low- and high-achieving children but the effects are not statistically significant. For children's effort displayed within the e-learning software (measured as the number of tasks solved voluntarily or the time spent on these tasks), our results seem to confirm the findings for motivation: there is a strong and significant heterogeneous treatment effect for low-achieving children (0.39 SD ($p = .044$) for number of tasks, 0.41 SD ($p = .078$) for time spent on voluntary practice); the linear combination with the treatment effect is, however, not significant.[9] Overall, results confirm that this type of relative performance feedback improved motivation and effort for low-achieving children without hurting middle- and high-achieving children.

Result 1: *The treatment strongly boosted motivation and effort for low-achieving children. For test outcomes, teacher ratings, and voluntary practice, low-achieving children showed large increases compared with the control group. Effects for self-rated motivation point in the same direction.*

Next, we ask whether these improvements in motivation and effort in the e-learning software for math actually translated into improved learning outcomes for math.[10] This is of crucial importance as one might be worried that the treatment might have increased motivation and effort—but only for playing a computerized "game" that has no relation to actual educational outcomes. However, this seems not to be the case in our setting. In the right-hand columns of Table 1, we report the results for our two math tests, measured after the treatment. The subgroup of low-achievers displays a strong and significant heterogeneous treatment effect on Math Multiplication/Division (0.74 SD, $p = .011$). The difference

[9] Yet, coefficients are substantial in size with up to 0.34 SD. Note that for this outcome, we are not able to control for baseline values; thus, estimations are presumably more noisy.

[10] Note that the task used to measure the *learning outcome* was different from the tasks used during the *treatment period* (math training); cf. Section A.3 in the SOM for details. Children received no feedback during evaluation waves and performance did not affect ranking as children could not score points in evaluation tasks.

Table 1: Effects of Relative Performance Feedback on Motivation, Effort, and Math Perf.

	Motiv. Task	Teac-r Mot.	Child-r Mot.	Vol. Tasks	Vol. Time	Math Add/Sub	Math Mult/Div
Treat × Low	0.469**	0.428***	0.258	0.388**	0.411*	0.364	0.742**
	(0.221)	(0.160)	(0.248)	(0.192)	(0.232)	(0.306)	(0.289)
Treat (Mid)	-0.132	0.046	-0.064	-0.179	-0.069	-0.250	-0.309
	(0.155)	(0.231)	(0.214)	(0.175)	(0.179)	(0.209)	(0.210)
Treat × High	0.233	0.030	0.249	-0.228	-0.157	0.365	0.399
	(0.213)	(0.223)	(0.272)	(0.256)	(0.228)	(0.260)	(0.250)
School FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Treat +	0.337**	0.474*	0.194	0.209	0.342	0.113	0.433*
Treat × Low	(0.150)	(0.282)	(0.206)	(0.199)	(0.301)	(0.211)	(0.234)
N	378	378	361	378	378	378	378

Notes: OLS regressions with post-treatment level of the respective outcomes as regressand. "Treat + Treat × Low" refers to the linear combination of the coefficients for "Treatment" and the interaction of "Treatment" with "Low-Achievers"; it indicates the difference between low-achieving children in the treatment group and low-achieving children in the control group. Standard errors in parentheses are clustered at the class level and corrected for small number of clusters using biased-reduced linearization (BRL, Bell & McCaffrey 2002). * p<.10, ** p<.05, *** p<.01

between low-achievers in treatment and control group for this learning outcome is large and significant at the 10-percent level (0.43 SD, $p = .065$). For Math Addition/Subtraction, the coefficient is positive but not significant.[11] Apparently, high-achievers are not hurt by this type of relative performance feedback—if anything, they also improve their performance. For this specific outcome, effects of the treatment on middle-achievers point in a negative direction—however, similar to the high-achievers, none of the coefficients are statistically significant.

Result 2: *Improved motivation and effort for the low-achievers translated into improved learning outcomes in math. While effects on the Math Addition / Subtraction tasks are not statistically significant, improvements for the Math Multiplication / Division tasks are large, and low-achievers in the treatment group perform significantly better than low-achievers in the control group with improvements amounting to about 80% of the initial gap between low- and middle-achievers in this task.*

To quantify the size of our treatment effect, we can compare the improvement of the low-achieving children in the treatment group with the initial ability gap in Math Multiplication/Division between low-achieving and middle-achieving children. Prior to treatment, low-achievers perform on average 0.54 SD worse than the group of middle-achievers. Thus, by improving their scores on Math Multiplication/Division by 0.43 SD, our treatment closed about 80% of the gap between low- and middle-achieving children in this specific math task.

[11] The fact that improvements for low-achievers seem to transfer to actual improvements in educational outcomes is supported by findings on teacher-rated math abilities and math grades, see Section A.2 in the SOM.

Is this increase in math performance (and motivation) actually driven by the new type of relative performance feedback we use in this study? In order to investigate the channels through which our treatment worked, we tested the following hypothesis: Are effects of *relative* compared with *individual* performance feedback stronger for children who receive more positive feedback in terms of the amount of points earned? To avoid trifold interactions, we simply restrict the sample to children who earn an above-median number of points on days where only performance improvements matter (i.e., days when no *Sprint 1* occurs, see Figure S1 in the SOM). Conducting the same analyses as in Table 1 using the restricted sample (n = 183, including 67 low-achievers), all coefficients of the interaction term (treatment x low-achiever) become substantially larger and most are significant, despite the sample size being reduced by more than 50%. The linear combinations indicating the difference between low-achievers in the treatment group and those in the control group are large and significant: treated low-achievers who receive above-median levels of points perform 0.43 SD better on the motivation test ($p = .026$), are rated 0.44 SD higher on motivation by their teachers ($p = .082$), they practice more voluntarily (0.41 SD, $p = .042$), and they perform better in the Math Multiplication/Division task by 0.67 SD ($p = .027$)—compared with low-achievers who receive above-median levels of points in the control group (all results are reported in Table S7 in the SOM).

Trying to pin down the mechanism of the treatment even closer, we analyzed whether children who were able to improve their *rank* during the time of the intervention were also the children who improved most on the *outcome* measures. We computed the average rank[12] on days 1–7 and on days 8–15 of the training (see Figure 2) and restricted the sample to children who *improved* their average rank from the first half to the second half of the intervention (n = 163, including 54 low-achievers). Results are even more striking than for the analyses using points earned: low-achievers in the treatment group who improved their rank in the second half of the intervention compared with the first half are 0.74 SD better in the motivation test ($p < .0001$), are rated 0.61 SD ($p = .045$) higher on motivation by their teachers, do 0.48 SD ($p = .035$) more voluntary practice tasks, and perform 0.71 SD ($p = .020$) better in the Math Multiplication/Division task (all results reported in Table S8 in the SOM).

Importantly, both these findings (i.e., for more points earned and rank improvement) cannot be interpreted as a causal effect because the number of points earned as well as the ranking are likely to be endogenous to the treatment condition.Despite this, we believe that

[12] Note that we can compute the (theoretical) rank for both treatment as well as control group children; yet, children in the control group never saw their actual rank. However, by using this theoretical rank we can create the perfect "control group within the control group" for this specific analysis.

Table 2: Effects of Relative Performance Feedback on Child-rated Outcomes

	Perc. Stress	Somatic Probl.	Self-efficacy	Liking Competit.
Treat × Low	0.271	-0.122	0.451	0.518***
	(0.188)	(0.246)	(0.280)	(0.178)
Treat (Mid)	0.086	0.037	-0.108	-0.246**
	(0.184)	(0.189)	(0.132)	(0.101)
Treat × High	-0.209	-0.022	0.070	0.137
	(0.295)	(0.191)	(0.168)	(0.169)
School FEs	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Treat +	0.357***	-0.085	0.343	0.273*
Treat × Low	(0.130)	(0.140)	(0.240)	(0.158)
N	363	360	360	360

Notes: OLS regressions with post-treatment level of the respective outcome as regressand. "Treat + Treat + Treat × Low" refers to the linear combination of the coefficients for "Treatment" and the interaction of "Treatment" with "Low-Achievers"; it indicates the difference between low-achieving children in the treatment group and low-achieving children in the control group. Standard errors in parentheses are clustered at the class level and corrected for a small number of clusters using biased-reduced linearization (BRL, Bell & McCaffrey 2002). * p<.10, ** p<.05, *** p<.01

our findings offer strong supportive evidence that this type of relative performance feedback is potentially very motivating and beneficial for low-achievers.

We also asked children to rate themselves on a number of dimensions (see Section A.3 in the SOM for details). In contrast to teachers, children can be viewed as blind to treatment.[13] We asked children how much stress they felt, whether they had somatic problems, how much self-efficacy they felt with respect to math, and how much they generally liked to compete with others. Again, results demonstrated substantial heterogeneous treatment effects for the low-achieving children (see Table 2). The subgroup of low-achieving children differed significantly between treatment and control group: children in the treatment group perceived more stress (0.36 SD, $p = .006$); at the same time, they did not report more somatic problems. Moreover, with respect to self-efficacy in the area of math, both the heterogeneous treatment effect for low-achievers as well as the difference between treatment and control group for low-achieving children was substantial but only close to statistical significance (0.45 SD with $p = .109$ and 0.34 SD with $p = .154$, respectively). When asked how much they liked to compete with others (in general, not specifically using this software), low-achieving children in the treatment group agreed much more strongly than low-achieving children in the control group. The heterogeneous treatment effect amounted to 0.52 SD ($p = .004$); the difference between low-achievers in treatment and control group was 0.27 SD ($p = .085$). For the attitude toward competition we also report a significant negative treatment effect on middle-achievers (0.25 SD, $p = .016$). Taken together, results from the child questionnaire

[13] Clearly, children were aware that they participated in a study but they did not know about the different treatment conditions and, thus, it is unlikely that their ratings were biased because they would have preferred the control condition or vice versa.

emphasized the heterogeneity of treatment effects for low-achieving children—however, they also point to a potential cost of relative performance feedback in terms of perceived stress (although this does not translate into somatic problems, at least not in the short run).

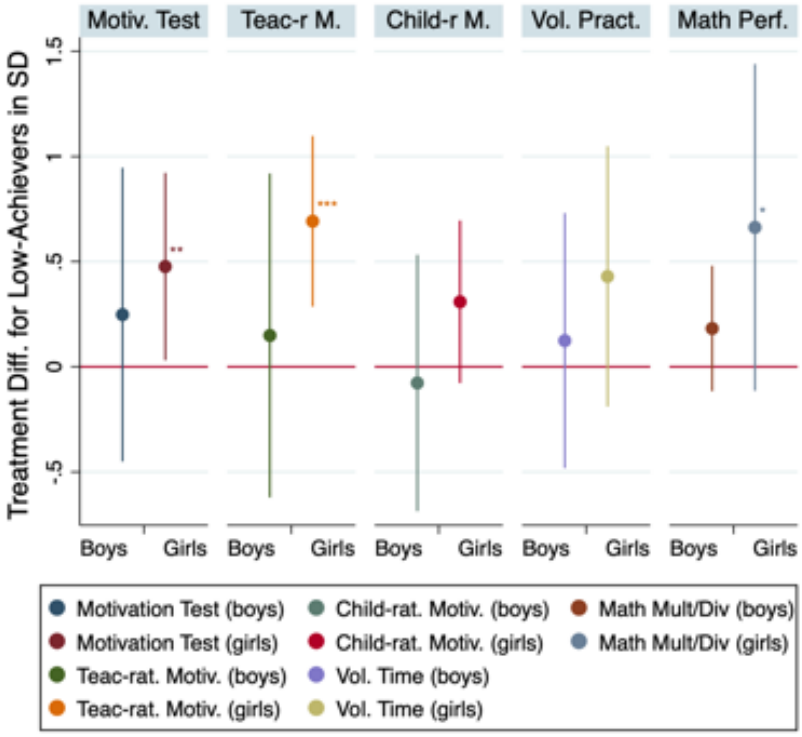
Result 3: *Low-achieving children reported higher perceived stress and displayed higher liking of competition. The increase in perceived stress does not translate into somatic problems. Also, we find suggestive evidence for higher self-efficacy among low-achievers.*

Previous studies have documented gender differences in response to relative performance feedback, mostly indicating stronger reactions for female subjects (Megalokonomou & Goulas, 2018; Azmat et al., 2019). At the same, our treatment introduced competition-like elements and a longstanding literature (see Niederle, 2016, for an overview) demonstrates substantial gender differences in competitiveness, already in childhood. For example, in a setting with children (in a similar age range as the present study) who learn their relative performance, Gneezy and Rustichini (2004) show that boys improved their performance on a non-incentivized task in a competitive setting (compared to a non-competitive setting). Thus, based on this literature one could expect that the improvements we documented in Results 1–3 would be stronger for boys. Therefore, it is of great interest whether there are gender differences in the reaction to the new type of relative performance feedback studied in the present paper. We find the following:

Result 4: *There are strong gender differences in the effect of the treatment on low-achievers. Within the subgroup of low-achieving children, mostly girls show improved motivation and increased math performance, but also report higher perceived stress. Low-achieving boys, in contrast, increase on self-rated self-efficacy in math and the liking of competition.*

We report findings for gender differences in treatment effects on motivation, effort, and math performance for low-achieving children in Figure 3. Because there were only 60 boys and 39 girls in the group of low-achieving children, we lost considerable power compared with our main results. However, we still report statistically significant differences between low-achieving children and their peers when looking at boys and girls separately. Specifically, treatment differences for low-achieving children in motivation seem mainly driven by girls (see Figure 3). Both for the motivation test as well as teacher-rated motivation, the difference between low-achieving children in the treatment group and low-achieving children in the control group is significant for girls (0.48 SD, $p = .037$; 0.69 SD, $p = .002$) and smaller and insignificant for boys. Child-rated motivation displayed a similar pattern, but the treatment difference for girls misses statistical significance (0.31 SD, $p = 0.111$). Similarly, differences

Figure 3: Gender Differences in Treatment Effects on Motivation and Math Performance



Notes: The figure is based on OLS regressions from Table 1 with post-treatment level of the respective outcomes as regressand, estimated separately for boys and girls. Within the group of low-achieving children, there are $n = 60$ boys and $n = 39$ girls. Dots show the linear combination of the coefficients for "Treatment" and the interaction of "Treatment" with "Low-Achievers"; they thus indicate the difference between low-achieving boys/girls in the treatment group and low-achieving boys/girls in the control group for each respective outcome (measured in SD of the outcome). Error bars show 95%-confidence intervals based on standard errors clustered at the class level. * p<.10, ** p<.05, *** p<.01

in effort measured as the time spent for voluntary practice tasks are mainly found for girls but do not reach significance (0.43 SD, $p = 0.163$). However, increased motivation for low-achieving girls apparently translated into strong and significant improvements for Math Multiplication/Division (0.66 SD, $p = .090$).

We also analyzed child-rated outcomes from Table 2 for gender differences (see Figure S4 in the SOM). Low-achieving girls were also those who reported higher stress levels (0.44 SD, $p = .007$). In contrast, low-achieving boys reported higher self-efficacy in math but the effect remains insignificant (0.68 SD, $p = .101$); they also stated a more positive attitude toward competition (0.58 SD, $p = .015$).

Overall, these findings point to important gender differences in how (low-achieving) children react to relative performance feedback of this type. The fact that we see strong behavioral changes for girls but not (or less strong) for boys relates to recent findings that providing information and relative performance feedback might diminish the gender gap in competitive environments (Ertac & Szentes, 2011; Wozniak, Harbaugh, & Mayr,

2016; Alan & Ertac, 2018), and could also, when combined with the fact that girls tend to underestimate their math abilities while boys tend to be overconfident (OECD, 2013), speak to the persistent gender gap observed for math abilities across many different education systems (see, for example, Guiso, Monte, Sapienza, & Zingales, 2008; OECD, 2014). Thus, further research should focus on analyzing this heterogeneity with respect to gender more closely.

Robustness Checks

To further support our findings, here we discuss some potential threats to the validity of our results. First, one could be worried that low-achieving children were especially motivated and outperformed their peers because the middle- and high-achieving children somehow "gave up" on the e-learning software. To check this, we plotted the performance of these three subgroups of children (low-, middle-, and high-achieving children; grouped based on their grades at baseline) and compared their development over time during our intervention period. Results in Figure S5 in the SOM show that performance in the math tasks was closely linked to grades: (i) children with better grades outperformed their peers in all 11 modules played during the intervention, (ii) this sorting in absolute performance (i.e., time needed to solve the *Sprints*) remained surprisingly stable over time, and (iii) especially children with good grades continuously *improved* their performance and became faster within each module over time. Hence, we conclude that *all* children spent considerable effort during the intervention period and that improvements for low-achieving children were not driven by negative effects on the part of middle- and high-achieving children. Second, given that children knew that they were rewarded for improvements in performance, one could generally worry that children might have *strategized* by being intentionally slow in the first *Sprint*. To engage with this argument, we provide Figure S6 in the SOM in which we present the average time needed in the first *Sprint* of each module. Evidently, there is no systematic increase in average time needed for the first *Sprint* over time; therefore, it is extremely unlikely that children played strategically. In addition, we see that for performance in the first *Sprint* of each module, sorting between different subgroups of children remained stable over time and that there were no differences between treatment and control group with respect to development over time and sorting into subgroups.[14]

[14] There is a jump in time needed between modules 1 and 2, but both modules were played on the very first day in varying order (children could choose their preferred order). Moreover, as the absolute time needed for each module is hard to compare between modules, additional analyses confirmed that there were no systematic differences for the average *improvements* (measured as the difference between time in the first *Sprint* and the average time in *Sprints 2–4*) in time needed when comparing the different modules or the different subgroups of children.

Both figures already indicate a third and final concern discussed here. Is it possible that low-achieving children were more motivated, simply because they had the possibility to earn many more points than the rest of their peers? In other words, did the calibration of points for the feedback leave other children "without a chance"? In order to check for that, we provide Figures S7–S8. Recall that in the first *Sprint* of each module, children were ranked according to their absolute performance within class (earning 1–10 points). In consequence, we see in Figure S7 that the ranking for the first *Sprint* closely mirrors the distribution of grades. For *Sprints 2–4*, however, points were distributed based on individual improvements (relative to average prior performance in this module). Accordingly, in the second *Sprint* the distribution of points more or less flipped around. Yet, already in *Sprint 3* the picture was less clear and in *Sprint 4* children seemed to earn points rather independent of their initial grade. To see how the ranking positions might have evolved, we also present cumulated points for each module in Figure S8—apparently, in *Sprints 3–4* children's cumulated number of points was more or less independent of their grade prior to the treatment. We interpret Figures S7–S8 as supportive evidence that our calibration of the feedback system was successful in (i) achieving ranking positions relatively independent of prior ability and (ii) in allowing for dynamic development of ranking positions over time. Overall, as our main results reveal, our new type of relative performance feedback was able to increase motivation and performance for low-achieving children without negatively affecting middle- and high-achieving children.

To further corroborate the robustness of our results, we also estimated our models excluding additional control variables, namely gender, age, and teacher-rated grades in Math and German prior to the treatment. Results of these analyses can be found in the SOM, Tables S9–S10. Qualitatively, all our findings are robust to excluding additional control variables; however, some coefficients are no longer statistically significant because standard errors become somewhat larger (loss in precision), and effects sizes become somewhat smaller, indicating non-perfect balance with respect to the control variables excluded here.[15]

4 Discussion

Management scientists, behavioral economists, and psychologists have devoted much attention (i) to the design of relative performance feedback systems, and (ii) to the effects of such social comparison information on individual choices and behavior. Results are in

[15] It is not surprising that balance was not perfect, given that we could only randomize between 20 classes. See Section 2.3 for further details on randomization.

general inconclusive; yet, many studies find detrimental effects for low-achievers. However, in the field of empirical educational research there is a clear lack of evidence on the various effects of relative performance feedback, especially with regard to classroom settings. This is remarkable, because gamified e-learning software is increasingly used even in early educational stages and because potential negative effects of relative performance feedback on learning outcomes of low-achievers would exacerbate educational inequalities.

Our results show that relative performance feedback does not always entail negative effects for low-achievers. For low-achieving children our treatment improves motivation, effort, and performance on math tasks. In turn, we do not find significant negative effects on middle- and high-achievers. An important question is why the gains for low-achievers on math performance are large and statistically significant only for Math Multiplication/Division, but not for Math Addition/Subtraction. A potential explanation is based on the fact that the e-learning software was designed to practice content in math which children learned towards the end of second grade (in contrast to teaching new content, see Section 2.4). While multiplication and division is still actively taught in the beginning of third grade, addition and subtraction (of numbers up to 100) is more or less settled. Thus, even for low-achievers there might be more opportunities to improve their performance in Math Multiplication/Division than in Math Addition/Subtraction.

In order to capture any negative effects of the treatment on other domains that are usually considered important in classroom environments, we also carefully measured children's risk-taking and social behavior with a set of outcomes (see SOM, Section A.3). Confirming the teachers' impressions (see Table S12), there were no negative effects of the treatment on risk-taking or social behavior (see SOM, Section A.2 and Table S11 for details). Yet, child-rated questionnaire items indicate that low-achieving children perceive the relative performance feedback as more stressful compared with the individual feedback, which would be a serious concern in educational applications. At the same time, low-achievers report more confidence in math and show a more positive attitude to competition—effects that one would generally evaluate very positively in school-related settings.

In addition to our main findings, we would like to highlight the importance of our results on gender differences for low-achieving children. In line with the findings by Gneezy and Rustichini (2004), we find a more positive attitude toward competition for boys than for girls in the child questionnaire at baseline. Similarly, we observe a strong gender gap in math performance in our baseline results (see Ellison & Swanson, 2010; Fryer & Levitt,

2010). If we look at the gender differences for low-achieving children, we see that our data reveal clear differences in responses by girls and boys. Figure 3 reveals that most treatment effects on actual behavior were driven by girls. These findings are in line with stronger reactions for females to relative performance feedback documented in Megalokonomou and Goulas (2018) and Azmat et al. (2019, although in this study, gender differences were not significant). Also, taking into account the gender differences for our self-reported measures, the evidence is very suggestive for boys being overconfident, as they are the ones to indicate higher levels of self-efficacy and liking of competition after the intervention (see Figure S4), while not (substantially) improving on any behavioral measure such as motivation or learning outcomes (see Figure 3).

Clearly, in our study gender differences are limited to the specific type of relative performance feedback we provide, and, importantly, to the specific subject used in the e-learning software, namely mathematics. This difference in the subject (or, the task) might also explain the differences in our results compared to Gneezy and Rustichini (2004) (they show that *boys* could improve in a running task in a competitive vs. non-competitive environment). Yet, in line with our findings, recent evidence shows that teaching the importance of effort in achievement (which is very much related to our *performance improvement* feedback) combined with performance feedback can eliminate the gender gap in a mathematics task for elementary school children (Alan & Ertac, 2018). Hence, as we have already noted when reporting our results, our findings could also partially help in explaining and closing the persistent gender gap observed for math abilities across many different countries.

More generally, our treatment effect *size* should be a lower bound of the true treatment effect of *relative* vs. *individual* performance feedback because, naturally, children in the control group also had *some* information about their relative performance (for example, by talking to their peers or occasionally seeing other children's screens). Similarly, if children in the control group (partially) understood the scoring mechanism, they also received "relative" feedback because they could infer from a high number of points that they did better than their peers (and vice versa).[16] Therefore, the control group also receives *some* relative performance feedback but, clearly, with a much lower intensity, frequency, and scope than in the treatment group.

Finally, based on our study design, there are a few limitations we would like to highlight. First, the results of our study have to be qualified as *short-term* findings because we measure all outcomes directly after the treatment. At the same time, the treatment itself must be

[16] It seems rather unlikely that children in third grade understood this complex scoring mechanism, including the differences between *Sprint 1* and *Sprints 2–4*. However, unfortunately we do not have any data on whether and how children actually understood the scoring mechanism.

qualified as being rather short and of low intensity compared with the actual use of e-learning software in educational practice, e.g., for a full school year. Recent evidence also suggests that relative performance feedback in an educational setting can have surprisingly long-lasting effects (although, for a college student sample, see Brade et al., 2018). Thus, the effects of using relative performance feedback over a longer period of time might be even stronger than in the present study. Likewise, we should note that long-run effects of feedback on performance improvements in itself have to be studied further because improvements on a given task are certainly limited (on the other hand, by regularly changing the task—similar to our different "modules" in the e-learning software—this effect could be circumvented).

Second, owing to our class-wise randomized design, our statistical power is rather low. Although we still report a number of statistically significant findings, standard errors are comparatively large and we cannot rule out that true effects sizes are substantially smaller (or larger). This is especially true for our analysis on gender differences for which sample size is smaller—still, results are statistically significant even in this small sample and display a consistent patter, thus, chance findings seem very unlikely. However, further studies using larger samples should shed more light on these gender differences in response to relative performance feedback on performance improvements. A related issue is that, although none of the effects on middle- and high-achieving children is statistically significant (except for the self-rated liking of competition for middle-achieving children, see Table 2), some coefficients are relatively large and indicate that there are potential effects of the treatment for which we do not have sufficient power to detect them. For example, high-achieving children in the treatment group seem to practice less on the voluntary tasks, and they also appear to slightly improve their math performance in both sets of math tasks (i.e., Addition/Subtraction and Multiplication/Division). Middle-achieving children, in turn, seem less motivated, show somewhat decreased math performance, and also report less liking of competition than middle-achieving children in the control group; hence, we need more research in order to analyze how this type of relative performance feedback affects outcomes along the ability distribution.

Third, our findings have to be qualified not only with respect to the *type* of relative performance feedback but also regarding the *comparison group* for which children receive relative performance information. Children in a classroom know the other children who are in front of (or behind) them, and this might make a difference for how they assess the information provided in the ranking. In fact, children might even have a prior about

other children's general math ability in school and might contrast this with the information they see in the ranking displayed in the e-learning software. While we argue that this is a very natural setting, especially when using e-learning software in the education system, it clearly constitutes a major difference to other studies using relative performance feedback, for example, with nationwide rankings for high school students (Megalokonomou & Goulas, 2018).

Fourth, because we randomized on class-level, a potential concern is that the effects of teachers are confounded with the treatment. Thus, varying attitudes of teachers toward the treatment could clearly influence how they evaluate children and could, potentially, also influence their behavior, e.g., how and what they teach. While we fully acknowledge this concern with respect to the teacher ratings, we argue that most of our measures come from within the e-learning software and could not be influenced by the teacher. The fact that we find large overlaps between the "objective" measures from within the e-learning software and the subjective, non-blind measures from the teacher questionnaire makes us confident that the results of our study capture actual effects of our treatment and are relevant to more general "class-room behavior", not only those specific to the outcome tasks used in the software.

5 Conclusion

Overall, our study provides first important evidence for the causal impact of a novel type of relative performance feedback for children in primary school age across the ability distribution. We conducted a randomized-controlled trial with about 400 children in regular teaching lessons and children used an e-learning software providing repetitive and continuous feedback on performance improvements over a period of five weeks. Our experimental design isolates the causal effect of *relative* vs. *individual* performance feedback on performance improvements on a broad range of outcomes using highly-standardized, computer-based tests and child- as well as teacher-questionnaires.

We report strong reactions of low-achieving children in the treatment group in terms of improved motivation and effort. Our evidence suggests that improved motivation and effort transferred to actual learning outcomes (i.e., math performance). The importance of the type of feedback for our findings is highlighted by the fact that children receiving more points and children who can improve their (average) ranking show the strongest improvements. Low-achieving children indicated higher levels of perceived stress but also increased self-efficacy in math and a more positive attitude toward competition. Finally,

even with our relatively small sample, we report strong gender differences in the reaction of low-achieving children to this type of relative performance feedback in mathematics. Specifically, improvements in motivation, effort, and math performance seem mainly driven by girls; boys indicate higher self-efficacy and a more positive attitude toward competition.

Thus, future research should focus on the effects of relative performance feedback that accounts for performance *improvements* in contrast to absolute performance levels. Similarly, gender differences in the reaction to relative performance feedback, especially for (low-achieving) children, deserve further investigation. On the application side, considering this type of relative performance feedback seems to be a promising path to reducing inequalities in education—both with respect to low-achievers as well as to the gender gap in mathematics. In light of the fast-growing use of e-learning software, already used in early educational stages, it is of key importance to consider the evidence on relative performance feedback when designing rankings, leader boards, etc., as motivational components of these applications. More broadly, the use of feedback on performance improvements could also be a useful tool to foster motivation and effort in workplace environments in which relative feedback on absolute performance can have negative effects at the lower end of the ability distribution (e.g., settings with very experienced and inexperienced workers). Further research should also focus on the dynamics and longer-term consequences of relative performance feedback on performance improvements.

Acknowledgments

We would like to thank Steffen Altmann, Alexander Cappelen, Gordon Dahl, Mathias Ekström, Michael Lovenheim, Erik Sørensen, Bertil Tungodden, and Lise Vesterlund for very helpful feedback. Moreover, we thank participants at the IPP Ideas Crunch Mainz, the 2nd IZA Workshop on Economics of Education, the MKWI Conference 2016, the DIGSSCORE+FAIR seminar, the IMEBESS 2018, and the ESA AP Meeting 2019 for helpful comments and feedback. Ann Thorhauer, Lion Huthmann, Tobias Metz, and Dirk Schweim provided great support to the project with their technical expertise in developing the e-learning software.

Funding

We gratefully acknowledge funding for the study by the German Federal Ministry of Education and Research (BMBF; Innovation and Technology Analysis, ITA2014, project no. PLI1643). Henning Hermes acknowledges partial financial support from the Research Council of Norway through its Centers of Excellence Scheme, project nos. 262675 (FAIR), 262636, & 250170F10. None of the sponsors had any involvement in study design, data analysis, writing of the paper, or decision to submit the article for publication.

Data Availability Statement

This paper uses proprietary data collected by the authors. The dataset is confidential because it combines a large set of preferences, attitudes, and skills of young school children; thus, it cannot be posted in a public repository. Researchers interested in replicating our results can obtain the data after signing a short research agreement with the authors of this study (send an email to henning.hermes@nhh.no to request further information). All our do-files are available in the Online Appendix.

Declaration of Interest

The authors declare no conflicts of interest.

IRB

The study received ethical approval from the joint Ethics Committee of the Faculty of Economics and Business Administration of the Goethe University Frankfurt (GU) and the Gutenberg School of Management & Economics of the Faculty of Law, Management and Economics of the Johannes Gutenberg University Mainz (JGU). The approval can be requested (search under the authors' names) at dekanat-fb03@uni-mainz.de.

References

Alan, S. & Ertac, S. (2018). Mitigating the Gender Gap in the Willingness to Compete: Evidence from a Randomized Field Experiment. *Journal of the European Economic Association*. doi:10.1093/jeea/jvy036

Ashraf, N., Bandiera, O., & Lee, S. S. (2014). Awards unbundled: Evidence from a natural field experiment. *Journal of Economic Behavior and Organization*, *100*(2014), 44–63. doi:10.1016/j.jebo.2014.01.001

Azmat, G., Bagues, M., Cabrales, A., & Iriberri, N. (2019). What you don't know... can't hurt you? a natural field experiment on relative performance feedback in higher education. *Management Science*, *forthcoming*. doi:10.1287/mnsc.2018.3131

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. doi:10.1016/j.jpubeco.2010.04.001

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. doi:10.1016/j.labeco.2015.02.002

Barankay, I. (2012). Rank incentives: Evidence from a randomized workplace experiment. *Business Economics and Public Policy Papers University, Working Paper, University of Pennsylvania*.

Bell, R. M. & McCaffrey, D. F. (2002). Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples. *Survey Methodology*, *28*(2), 169–181.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, Reciprocity, and Social History. doi:10.1006/game.1995.1027

Blanes i Vidal, J. & Nossol, M. (2011). Tournaments without prizes: evidence from personnel records. *Management Science*, *57*(10), 1721–1736. doi:10.1287/mnsc.1110.1383

Bolton, G. E. & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, *90*(1), 166–193. doi:10.1257/aer.90.1.166. arXiv: arXiv:1011.1669v3

Brade, R., Himmler, O., & Jäckle, R. (2018). *Normatively Framed Relative Performance Feedback – Field Experiment and Replication* (Working Paper No. 88830). MPRA.

Brandstätter, V. (2005). Tests und Tools - Der objektive Leistungs-motivations-Test (OLMT). *Zeitschrift für Personalpsychologie*, *4*(3), 132–137. doi:10.1026/1617-6391.6.3.129

Bursztyn, L. & Jensen, R. (2015). How Does Peer Pressure Affect Educational Investments? *The Quarterly Journal of Economics*, *130*, 1329–1367. doi:10.1093/qje/qjv021.Advance

Card, D., Mas, A., Moretti, E., & Saez, E. (2012). Inequility at work: The effec of peer salaries on job satisfaction. *American Economic Review*, *102*(6), 2981–3003. doi:10.1257/aer.102.6.2981

Charness, G., Masclet, D., & Villeval, M. C. (2014). The Dark Side of Competition for Status. *Management Science*, *60*(1), 38–55.

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A Focus Theory of Normative Conduct: Recycling the Concept of Norms to Reduce Littering in Public Places. *Journal of personality and social psychology*, *58*(6), 1015–1026.

Claro, S., Paunesku, D., & Dweck, C. S. (2016). Growth mindset tempers the effects of poverty on academic achievement. *Proceedings of the National Academy of Sciences*, *113*(31), 8664–8668.

Coleman, J. (1990). *Foundations of Social Theory.* Cambridge, MA: Belknap Press.

Connelly, B. L., Tihanyi, L., Crook, T. R., & Gangloff, K. a. (2014). Tournament Theory: Thirty Years of Contests and Competitions. *Journal of Management*, *40*(1), 16–47. doi:10.1177/0149206313498902

Delfgaauw, J., Dur, R., Sol, J., & Verbeke, W. (2013). Tournament Incentives in the Field: Gender Differences in the Workplace. *Journal of Labor Economics*, *31*(2), 305–326. doi:10.1086/667996

der Kleij, F. M. V., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research*, *85*(4), 475–511.

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From Game Design Elements to Gamefulness : Defining " Gamification ". In *Mindtrek '11 proceedings of the 15th international academic mindtrek conference: Envisioning future media environments* (pp. 9–15). Tampere, Finland: ACM.

Dijk, O., Holmen, M., & Kirchler, M. (2014). Rank Matters - The Impact of Social Competition on Portfolio Choice. *European Economic Review*, *66*(2014), 97–110.

Dobrescu, L., Faravelli, R., Megalokonomou, R., & Motta, A. (2019). *Rank incentives and social learning: Evidence from a randomized natural experiment*. Working Paper.

Ellison, G. & Swanson, A. (2010). The Gender Gap in Secondary School Mathematics at High Achievement Levels: Evidence from the American Mathematics Competitions. *Journal of Economic Perspectives*, *24*(2), 109–128. doi:10.1257/jep.24.2.109

Eriksson, T., Poulsen, A., & Villeval, M. C. (2009). Feedback and incentives: Experimental evidence. *Labour Economics*, *16*(6), 679–688. doi:10.1016/j.labeco.2009.08.006

Ertac, S. & Szentes, B. (2011). *The Effect of Information on Gender Differences in Competitiveness: Experimental Evidence* (Working Paper No. 1104). Koc University-TUSIAD Economic Research Forum.

Fehr, E. & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, *114*(3), 817–868.

Festinger, L. (1954). A Theory of Social Comparison Processes. *Human Relations*, *7*(2), 117–140.

Fischer, M. & Wagner, V. (2018). *Effects of Timing and Reference Frame of Feedback: Evidence from a Field Experiment* (Working Paper No. 1820). Gutenberg School of Management and Economics, Johannes Gutenberg-Universität Mainz.

Fryer, R. & Levitt, S. (2010). An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics*, *2*(2).

Gill, D., Kissova, Z., Lee, J., & Prowse, V. L. (2018). First-Place Loving and Last-Place Loathing: How Rank in the Distribution of Performance Affects Effort Provision. *Management Science, Articles in Advance*, *2018*, 1–14. doi:10.2139/ssrn.2641875

Gneezy, U., Niederle, M., & Rustichini, A. (2003). Performance in Competitive Environments: Gender Differences. *The Quarterly Journal of Economics*, *118*(3), 1049–1074. doi:10.1162/00335530360698496

Gneezy, U. & Rustichini, A. (2004). Gender and Competition at a Young Age. *American Economic Review*, *94*(2), 377–381. doi:10.1257/0002828041301821

Guiso, L., Monte, F., Sapienza, P., & Zingales, L. (2008). Culture, Gender, and Math. *Science*, *320*(5880), 1164–1165. doi:10.1126/science.1154094

Haenni, S. (2019). Ever tried. Ever failed. No matter? On the demotivational effect of losing in repeated competitions. *Games and Economic Behavior*. doi:10.1016/j.geb.2019.03.012

Hannan, R. L., Krishnan, R., & Newman, A. H. (2008). The effects of disseminating relative performance feedback in tournament and individual performance compensation plans. *Accounting Review*, *83*(4), 893–913. doi:10.2308/accr.2008.83.4.893

Hannan, R. L., McPhee, G. P., Newman, A. H., & Tafkov, I. D. (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. & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. doi:10.3102/003465430298487

Kahneman, B. D., Knetsch, J. L., & Thaler, R. (1986). Fairness as a Constraint on Profit Seeking: Entitlements in the Market. *The American Economic Review, 76*(4), 728–741.

Kirchler, M., Weitzel, U., & Lindner, F. (2018). Rankings and Risk-Taking in the Finance Industry. *Journal of Finance*, *in press*, 1–79. doi:10.2139/ssrn.2760637

Kuziemko, I., Buell, R. W., Reich, T., & Norton, M. I. (2014). "Last-Place Aversion": Evidence and Redistributive Implications. *The Quarterly Journal of Economics*, *129*(1), 105–149. doi:10.1093/qje/qjt035.Advance

Lazear, E. P. & Rosen, S. (1981). Rank-Order Tournaments as Optimum Labor Contracts. *Journal of Political Economy*, *89*(5), 841–864. doi:10.1086/261010

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., ... Brown, R. a. (2002). Evaluation of a Behavioral Measure of Risk Taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied*, *8*(2), 75–84.

Linde, J. & Sonnemans, J. (2012). Social comparison and risky choices. *Journal of Risk and Uncertainty*, *44*, 45–72. doi:10.1007/s11166-011-9135-z

Megalokonomou, R. & Goulas, S. (2018). *Knowing who you are - The Effect of Feedback Information on Short and Long Term Outcomes*. University of Warwick, Department of Economics.

Niederle, M. (2016). Gender. In *Handbook of Experimental Economics* (Volume 2). Princeton Univers. Press.

Niederle, M. & Vesterlund, L. (2007). Do Women Shy Away From Competition? Do Men Compete Too Much? *The Quarterly Journal of Economics*, *122*(3), 1067–1101. doi:10.1162/qjec.122.3.1067

OECD. (2013). *Pisa 2012 results: Ready to learn (volume iii)*. doi:https://doi.org/https://doi.org/10.1787/9789264201170-en

OECD. (2014). *Mathematics performance (PISA)*. OECD Publishing. doi:10.1787/04711c74-en

Prendergast, C. (1999). The Provision of Incentives in Firms. *Journal of Economic Literature*, *37*(1), 7–63. doi:10.1257/jel.37.1.7. arXiv: z0022

Sailer, M., Hense, J. U., Mayr, S. K., & Mandl, H. (2017). How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. *Computers in Human Behavior*, *69*(2017), 371–380. doi:10.1016/j.chb.2016.12.033. arXiv: arXiv:1011.1669v3

Schwarzer, R. & Jerusalem, M. (1999). *Skalen zur Erfassung von Lehrer- und Schülermerkmalen. Dokumentation der psychometrischen Verfahren im Rahmen der Wissenschaftlichen Begleitung des Modellversuchs Selbstwirksame Schulen*. Berlin. arXiv: 0208024 [gr-qc]

Slovic, P. (1966). Risk-Taking in Children: Age and Sex Differences. *Child Development*, *37*(1), 169–176.

Smither, R. D. & Houston, J. M. (1992). The Nature of Competitiveness: The Development and Validation of the Competitiveness Index. *Educational and Psychological Measurement*, *52*, 407–418.

Snoeren, F. & Hoefnagels, C. (2014). Measuring perceived social support and perceived stress among primary school children in the Netherlands. *Child Indicators Research*, *7*, 473–486. doi:10.1007/s12187-013-9200-z

Tran, A. & Zeckhauser, R. (2012). Rank as an inherent incentive: Evidence from a field experiment. *Journal of Public Economics*, *96*(10), 645–650. doi:10.1016/j.jpubeco.2012.05.004

Veblen, T. (1899). *The theory of the leisure class*. New York, NY: Macmillan.

Wilde, M., Bätz, K., Kovaleva, A., & Urhahne, D. (2009). Testing a short scale of intrinsic motivation. *Zeitschrift für Didaktik der Naturwissenschaften*, *15*, 31–45. arXiv: 2404161

Wozniak, D., Harbaugh, W. T., & Mayr, U. (2016). *The Effect of Feedback on Gender Differences in Competitive Choices.* (SSRN Scholarly Paper No. ID 1976073). Social Science Research Network. Rochester, NY.

Part

Supplementary Online Material

Table of Contents

A Supplementary Text

A.1 Descriptive Statistics

In Figures S2 and S3, we present histograms for all outcome measures prior to the treatment (W1) and after the treatment (W2). Most outcomes seem well-distributed and have sufficient variance; however, for math performance our outcome task suffers from ceiling effects. Because we were afraid that low-achieving children might also be the ones who still have "room to improve" on the math tasks, which could partially explain our heterogeneous results for this subgroup, we conducted the following robustness check: we identified all children with "room for improvement" and analyzed whether there is a heterogeneous treatment effect for this subgroup. Findings are presented in Table S13; there is no significant treatment effect for this subgroup. Given these results and in light of the fact that findings for motivation and teacher ratings also point in the same direction, we conclude that this ceiling effect for our measures of learning outcomes in math is not driving our findings.

A.2 Further Results

Table S11 also reports results for risk attitudes and social behavior—a set of outcomes for which previous studies found detrimental effects of relative performance feedback; many teachers also expressed concerns that the treatment might have negative effects in these domains. Therefore, we also wanted to evaluate the effects on these types of behavior and used several tasks to measure them (see Section A.3 for details on tasks). Using combined measures[17] for all tasks concerning risk and all tasks concerning social behavior, we report no effects of the treatment on either dimension along the ability distribution. Most coefficients are close to zero (if anything, low-achieving children take *fewer* risks but the effect is not statistically significant).

In addition to our test outcomes, we also asked teachers to rate the relevant outcomes for their children in class, both prior to treatment as well as post-treatment (see Section A.3 for details on measurement). Results for analyses using these ratings as an outcome are reported in Table S12. Two concerns make results on teacher ratings difficult to interpret. First, teachers were not blind to treatment, thus, their ratings might (partially) also reflect their attitude to the treatment (or the control group type of software). Second, as the treatment was randomized between classes, the treatment might have shifted the whole

[17] We combine the different measures into a single index because we want to account for measurement error and avoid problems with multiple testing. Effects are very similar when looking at the tasks one by one.

distribution within a class, which is hard to observe for the teacher (or hard to disentangle from a general time trend). Therefore, we are cautious in interpreting the results on teacher ratings; we do, however, report them to provide a complete and transparent picture of the effect of our treatment.

Generally, results for teacher-rated outcomes largely mirror our findings for test outcomes. While none of the treatment interactions is significant, teachers do rate low-achieving children in the treatment group better on math abilities than the comparable children in the control group. Results for grades in math point in the same direction but are insignificant. For risk attitudes and social behavior, there are no differences in teacher ratings between the two experimental conditions. Finally, for teacher-rated stress there is no significant difference between the treatment conditions. All in all, teachers largely confirm the results of our test outcomes, which supports the notion that the effects of this type of relative performance feedback do not only affect the outcome tasks we measure but, to some extent, seem to generalize to everyday classroom behavior.

A.3 Detailed Description of Outcome Measures

We applied several standardized data collection methods to measure the treatment effects on the outcome variables. On the one hand, we observed children's behavior during the math training phase.

Voluntary practicing in mathematics e-learning software. Each day, after having completed the compulsory part of the math training, children had the opportunity to do additional, voluntary math tasks. Here, we measured the number of voluntarily solved math tasks (Vol. Tasks) and children's voluntary training time (Vol. Time) to evaluate children's motivation to practice.

On the other hand, we measured outcomes in the two evaluation waves before and after the training phase. We used highly standardized elicitation methods with automated video-based instructions. In order to account for the young age of our subjects, we mainly implemented game-based elicitation methods with an age-appropriate interface design. Every task was explained beforehand to the children with a video sequence and an audio explanation. Importantly, these measures in the two evaluation waves were not linked to the point system in the math training phase and entailed neither relative nor private

performance feedback (but were incentivized, see Section 2.1). The following methods were used to evaluate our outcomes.

Math Addition/Subtraction & Multiplication/Division. To measure the treatment effects on learning outcomes, we implemented two sets of math tasks on basic arithmetic operations (Math Addition/Subtraction and Math Multiplication/Division). Each task consisted of 10 single-choice tasks with four response options. Children were informed that they should try to answer the tasks as fast as possible. Providing a wrong answer caused a waiting time. The sequence of the tasks was randomized to prevent cheating. However, every child had to answer the same tasks. The type of task was very different from those in the training phase. Thus, our outcome depicts untrained math performance. The number of correctly answered tasks was interpreted as math performance (see Figure S11).

Motivation Task (Brandstätter, 2005). The Objective Achievement Motivation Task (Objektiver Leistungsmotivationstest; OLMT) is a test to measure children's intrinsic motivation. Children were presented with a 4x4 grid with squares and circles. The task was to count the number of circles for each grid. Children had to solve as many grids as possible by filling in the respective number of circles in an input field on the screen within a time frame of 120 seconds. If a child put in the right answer, the next grid was displayed. Importantly, children could not earn coins in the OLMT, and thus the game was not incentivized. Children played three different rounds of the OLMT in sequential order and each round lasted for 120 seconds (round 1: doing your best, round 2: setting individual goals, round 3: competing against a fictitious other person). The total number of correctly solved grids was used to measure intrinsic motivation (see Figure S12).

Child Questionnaire. To elicit children's self-reported motivation, perceived stress, somatic problems, self-efficacy, and their liking of competition (i.e., a positive attitude toward competition), we implemented a computer-based questionnaire. Overall, children had to answer 20 questionnaire items in which they had to state whether they agreed or disagreed with a certain statement (see Table S1). All statements were read aloud and children listened to the instruction via their headphones. We implemented five age-appropriate and icon-based response options (Icon 1 = strongly disagree, Icon 2 = disagree, Icon 3 = neither agree/nor disagree, Icon 4 = agree, Icon 5 = strongly agree [from left to right], see Figure S19). While the statement was read aloud, the response options were disabled. This

guaranteed that all children carefully listened to the whole statement before entering their answer.

Index Risk Attitude. We combined three different measures of risk attitudes into an index for risk attitudes by summing the standardized scores from the following tasks. First, we conducted the Balloon Analogue Risk Task (BART, Lejuez et al., 2002). Here, children were sequentially presented with 10 balloons on the computer screen. They could earn coins by pumping up each balloon (with mouse clicks). Each click inflated the balloon incrementally and with each click one point was added to a contemporary account. If children decided to stop pumping up the balloon, the points in the temporary account were moved to a permanent save account. Otherwise, if participants reached a randomly determined maximum number of pumps, the balloon exploded. This resulted in a loss of all points accrued for this balloon(see Figure S13). Second, we used the "Devil's Task" (Slovic, 1966) to elicit risk behavior. Subjects were presented with an array of 10 closed treasure chests on their computer screen. They could earn coins by opening these treasure chests (with mouse clicks). They were informed that nine chests contained a reward (one point) and one box contained a "devil". Each click on one of the chests containing a reward added one point to a contemporary account. If children decided to stop opening chests, the points in the temporary account were moved to a permanent save account. Otherwise, if participants drew the devil, they lost all points accrued for this round. The game was played over seven rounds(see Figure S14). Third, we implemented a standard Lottery Task in which children received an endowment of 10 coins. They could decide how many coins to invest in a lottery. Children could move coins from the safe coin stack (on the left) to the lottery coin stack (on the right) by clicking on arrows. The lottery consisted of two cards: a loss card (sad smiley) and a win card (happy smiley). If subjects drew the loss card they lost all coins at stake. If they drew the win card the coins at stake were doubled(see Figure S15).

The index of risk attitudes was calculated as an unweighted average of the number of clicks on balloons in the BART, the number of opened chests in the Devil's Task, and the number of coins set in the lottery (all values standardized to mean = 0 and SD = 1).

Index Social Behavior. Similarly, to proxy social behavior (toward classmates), we computed an index comprised of standardized values of three different decisions children had to take. We conducted a Trust Game (Berg, Dickhaut, & McCabe, 1995) in which two children interacted as truster and trustee. To determine which children interacted, every

child was randomly and anonymously matched to a classmate. First, children had the truster role and they were informed that they played with a classmate (trustee) in their class. Children received an endowment of 10 coins. They had to decide how many of these 10 coins to send to their anonymous classmate. Each coin that the child entrusted to the other child was doubled. Further, children were informed that the trustee, in a second step, could send some or all of these received coins back to them(see Figure S16). In the receiving situation of the trust game, children had the trustee role. The trustee first saw the number of coins that had been entrusted to them (already doubled). Then they decided on how many coins to return to the anonymous classmate. Children could decide to send all, some, or no coins back to the truster. If the truster initially did not send any coins children got the information that their matched partner did not send any coins and the decision was skipped(see Figure S17).In addition, we used a standard Dictator Game (Kahneman, Knetsch, & Thaler, 1986) to measure children's willingness to share with an anonymous classmate. First, children received an initial endowment of 10 coins. Second, they had to decide how to split the 10 coins between themselves (dictator) and an anonymous child in class (receiver). Children were randomly and anonymously matched to a classmate(see Figure S18).

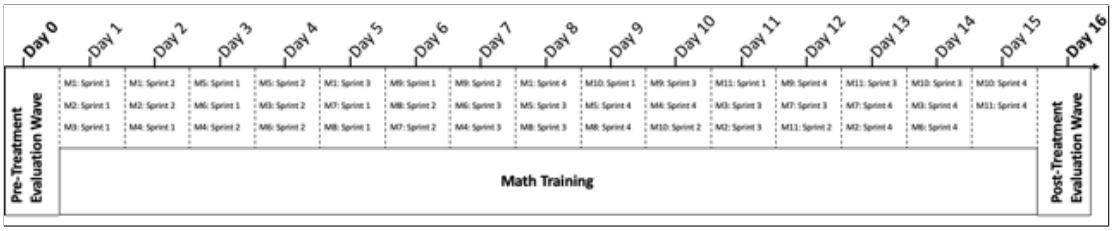
The index of social behavior was calculated as an unweighted average of the number sent in the Trust Game, the share returned in the Trust Game (if available), and the number sent in the Dictator Game (all values standardized to mean = 0 and SD = 1).

Teacher Questionnaire. Teachers received a paper-based questionnaire that they filled out during the two lessons in which evaluation waves took place (before and after the training phase). Within the questionnaire they were asked to state how much they agreed with a statement about every child in their class. In addition, we asked for each child's current grade in Math and German. Please refer to Table S2 for the questionnaire items.

B Supplementary Figures

Detailed Timeline of the Field Study

Figure S1: Timeline of the Field Study



Descriptive Statistics

Figure S2: Distribution of Outcome Variables prior to Treatment (W1)

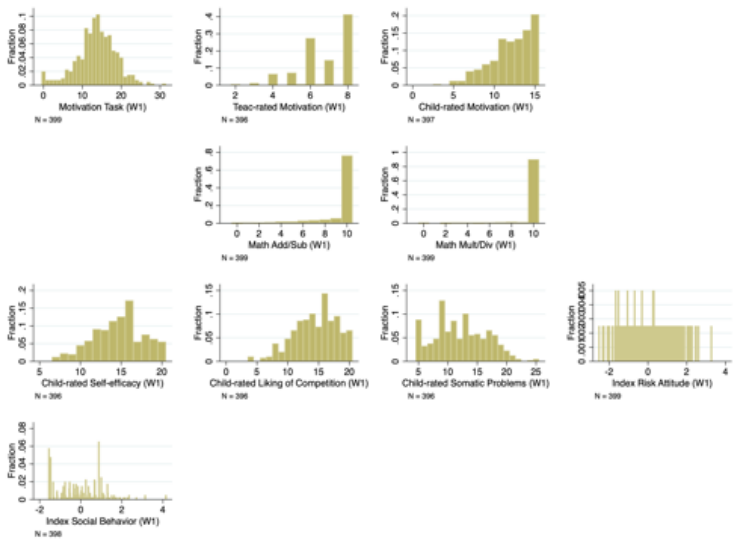
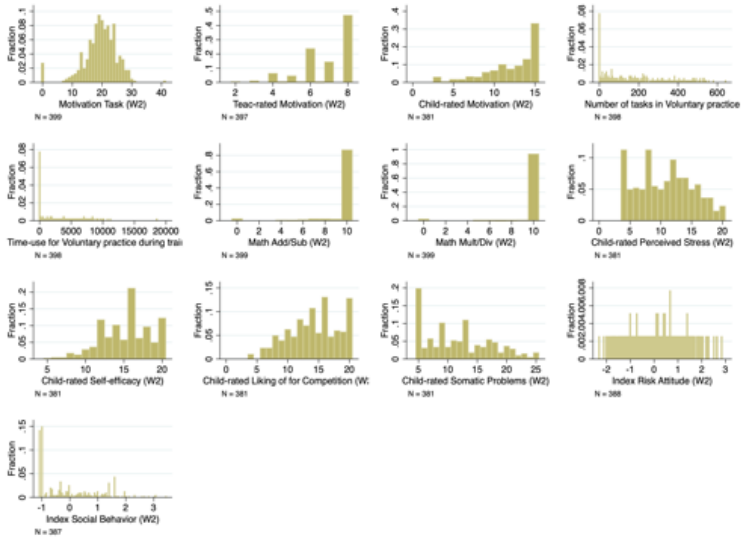
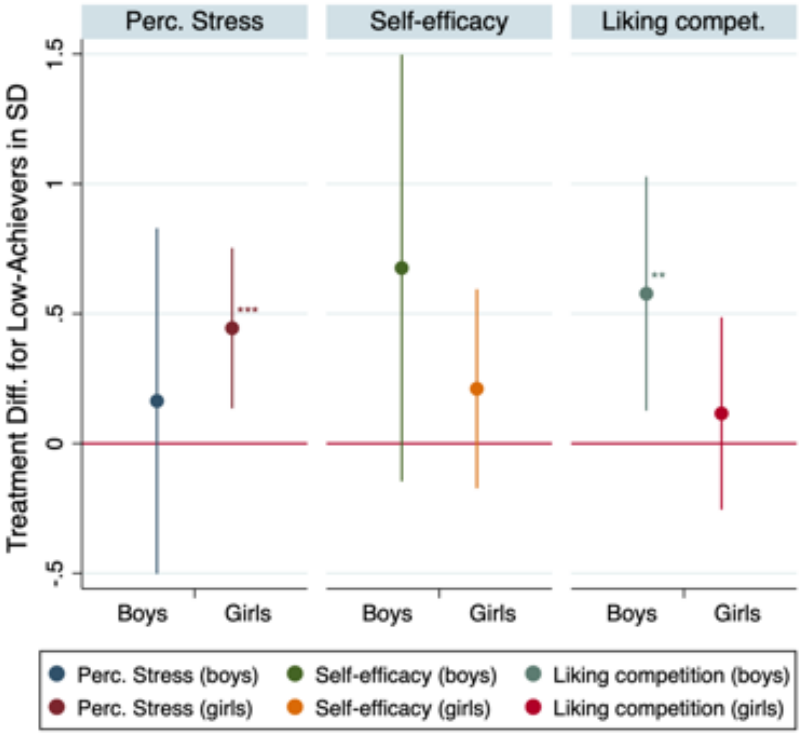


Figure S3: Distribution of Outcome Variables after Treatment (W2)



Gender Differences

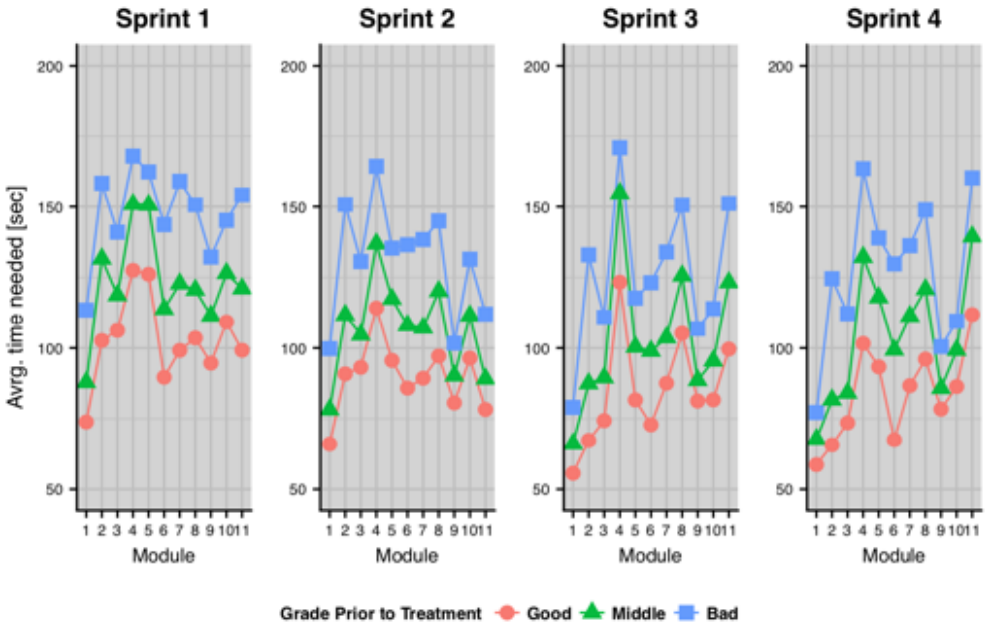
Figure S4: Gender Differences in Treatment Effects on Child-rated Outcomes



Notes: The figure is based on OLS regressions from Table 2 with post-treatment level of the respective outcomes as regressand, estimated separately for boys and girls. Within the group of low-achieving children, there are $n = 60$ boys and $n = 39$ girls. Dots show the linear combination of the coefficients for "Treatment" and the interaction of "Treatment" with "Low-Achievers"; they thus indicate the difference between low-achieving boys/girls in the treatment group and low-achieving boys/girls in the control group for each respective outcome (measured in SD of the outcome). Error bars show 95%-confidence intervals based on standard errors clustered at the class level. * p<.10, ** p<.05, *** p<.01

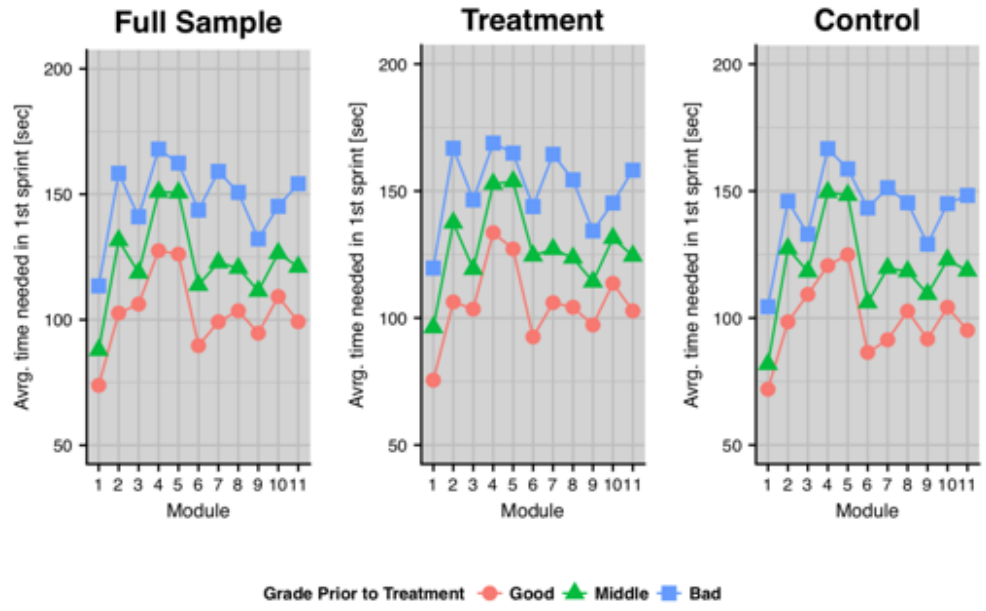
Performance over Modules in Math Software

Figure S5: Average Time Needed in *Sprints 1–4*, Children Grouped by Grade



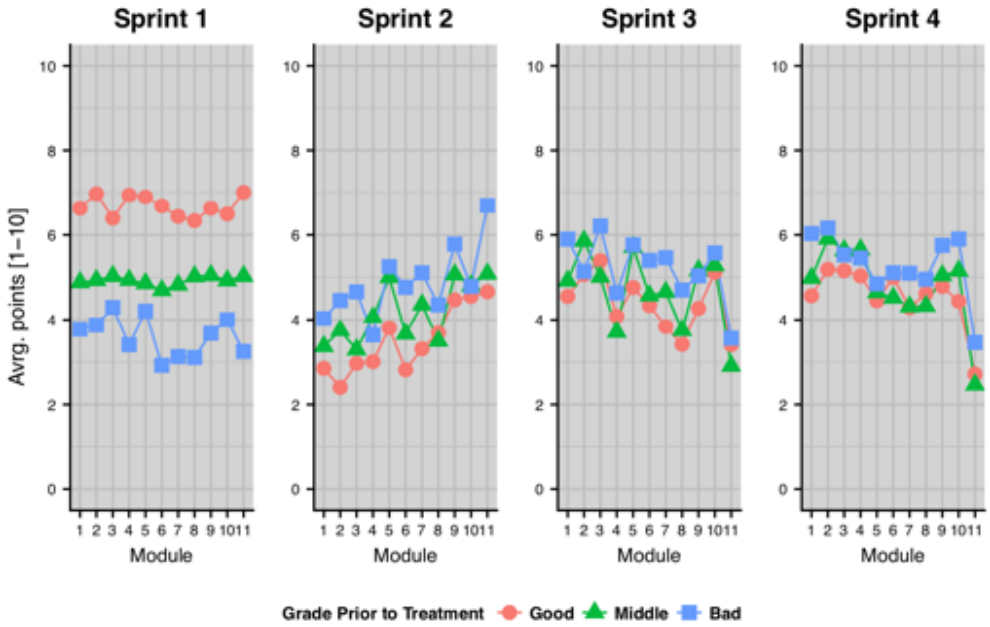
Notes: We classified children with a very good grade (= 1) as "Good" (N = 114). Children with a grade higher than 3 were classified as "Bad" (N = 99). All the others were classified as "Middle" (N = 165). Grades range from 1 (very good) to 6 (insufficient).

Figure S6: Average Time Needed in *Sprint 1*, Children Grouped by Grade and Treatment Status



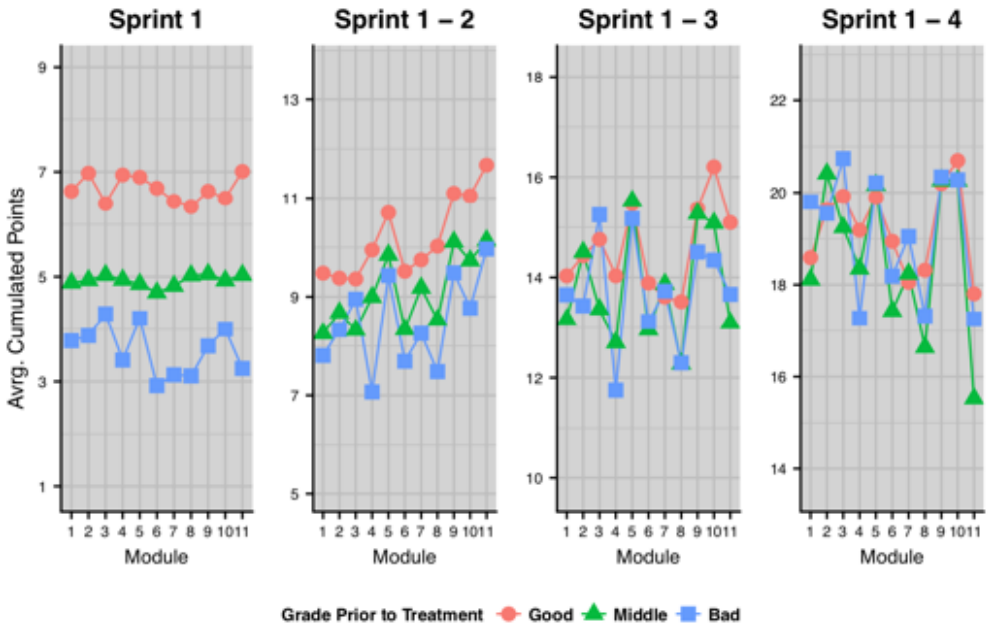
Notes: We classified children with a very good grade (= 1) as "Good" (N = 114). Children with a grade higher than 3 were classified as "Bad" (N = 99). All the others were classified as "Middle" (N = 165). Grades range from 1 (very good) to 6 (insufficient).

Figure S7: Points Received in *Sprints 1–4* for each Math Module, Children Grouped by Grade



Notes: We classified children with a very good grade (= 1) as "Good" (N = 114). Children with a grade higher than 3 were classified as "Bad" (N = 99). All the others were classified as "Middle" (N = 165). Grades range from 1 (very good) to 6 (insufficient).

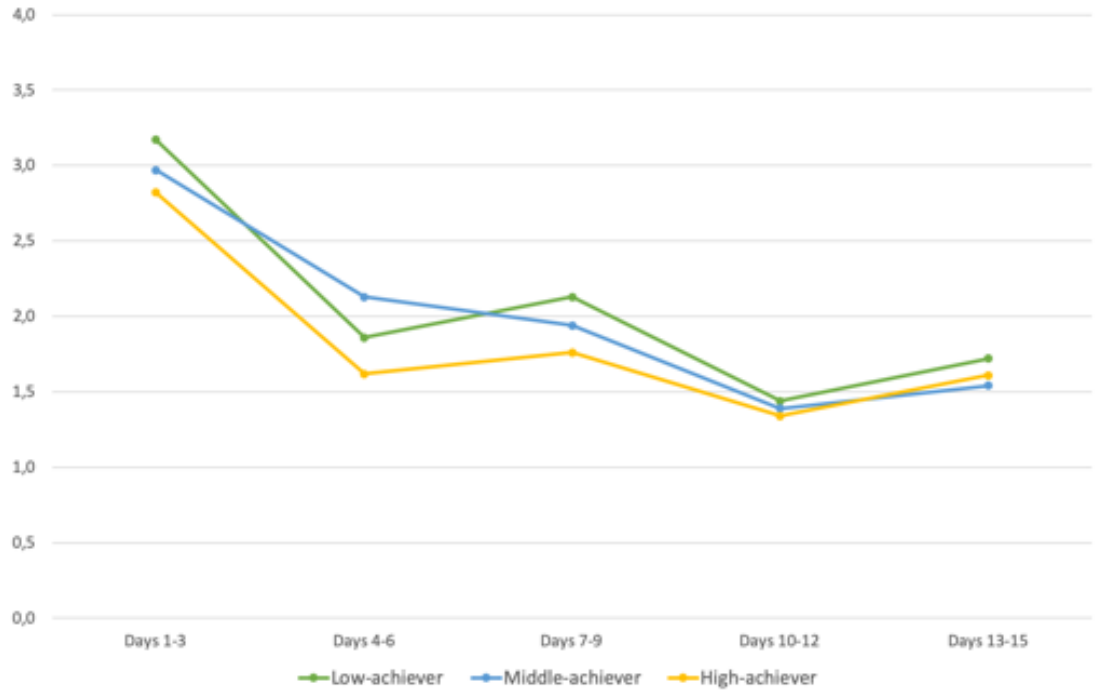
Figure S8: Accumulated Points Received in *Sprints 1–4* for each Math Module, Children Grouped by Grade



Notes: We classified children with a very good grade (= 1) as "Good" (N = 114). Children with a grade higher than 3 were classified as "Bad" (N = 99). All the others were classified as "Middle" (N = 165). Grades range from 1 (very good) to 6 (insufficient).

Development of Dynamics in Ranking over Time

Figure S9: Mean Standard Deviation of Average Rank over Time for the Treatment Group



Notes: To analyze whether the cumulative point system led to distortions in the dynamics of rankings, we plot the mean standard deviation of the average rank during a three day time window for the treatment group. Dots constitute the mean standard deviation of the average rank for low-, middle, high-achievers, using the standard deviation of a child's average rank across all modules conducted during the three days (i.e., mostly nine modules). The figure shows that, while the dynamics in ranking position clearly decreases from days 1–3 to days 13-15 (which is an inherent feature of the cumulative point system the ranking is based on), there is no substantial difference for dynamics in ranking across the three subgroups of children. Also, after the first three days, dynamics of the ranking remains fairly stable and even seems to slightly increase towards the end of the intervention period.

E-Learning Software

Figure S10: Math Modules in the E-Learning-Software

(a) M1 – Which number?

(b) M2 – Double and Half

(c) M3 – Decompose

(d) M4 – Mental Maths - Minus

(e) M5 – Count in Steps

(f) M6 – Times Table

(g) M7 – Multiplication Table

(h) M8 – Mental Maths - Plus

(i) M9 – Tag the Number

(j) M10 – Multiplication Table

(k) M11 – Number Wall

Figure S11: Screenshot of the Single-Choice Math Task

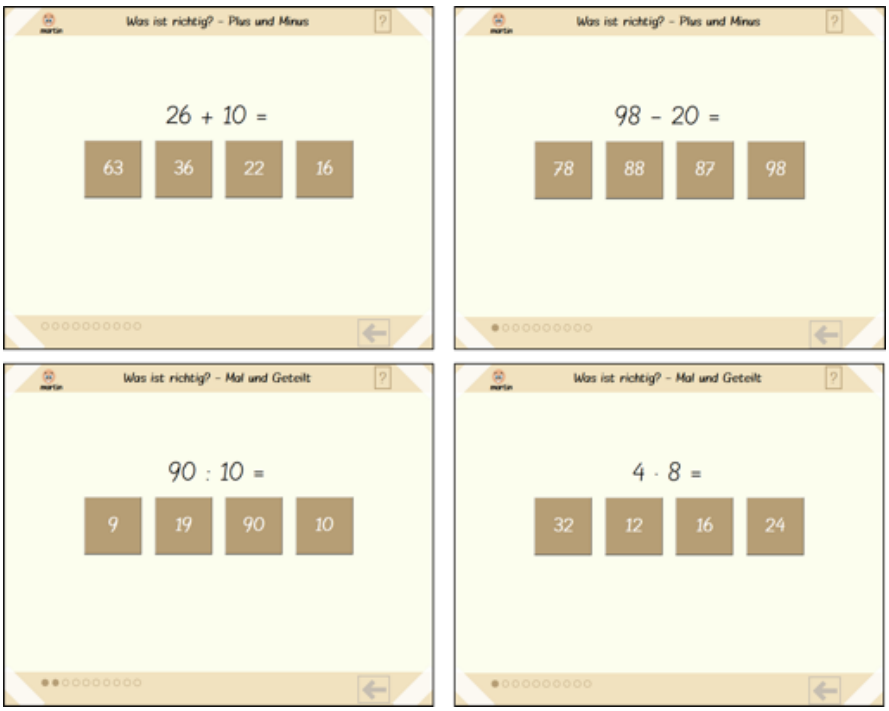


Figure S12: Screenshot of the Objective Achievement Motivation Test (OLMT)

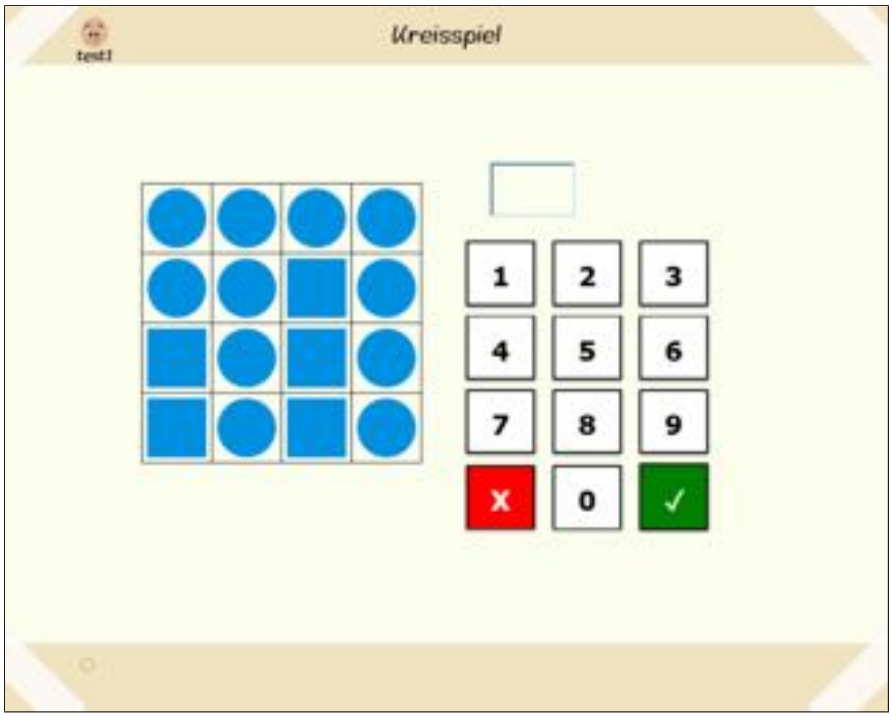


Figure S13: Screenshot of the Balloon Analogue Risk Task (BART)

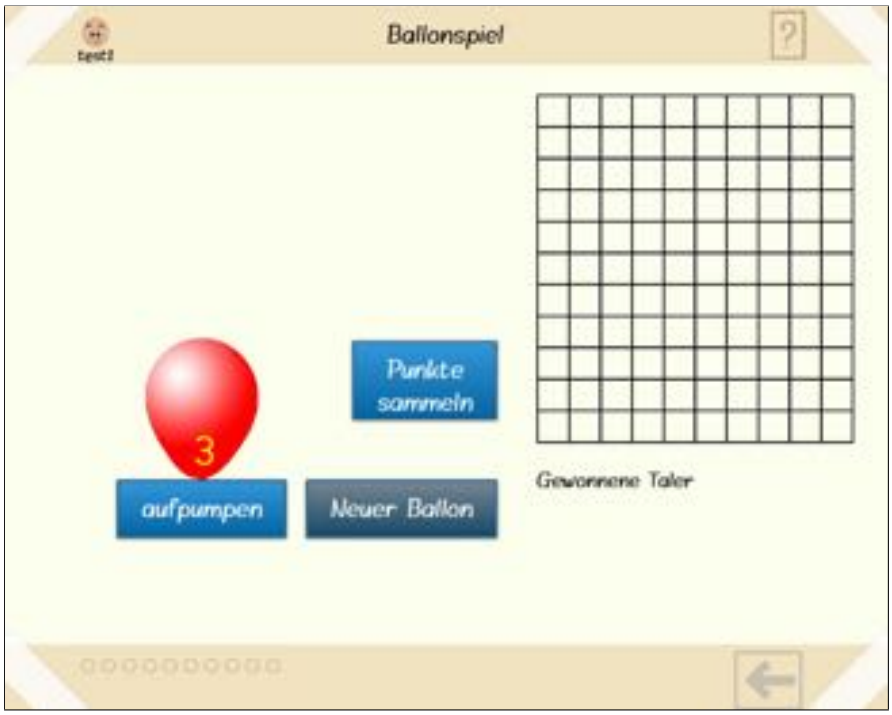


Figure S14: Screenshot of the Devil's Task

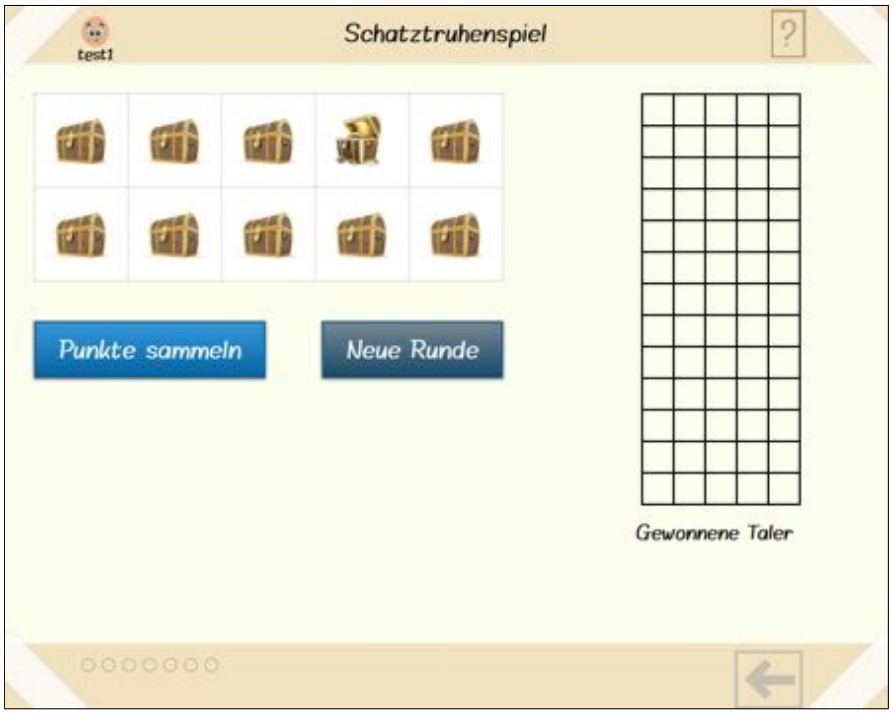


Figure S15: Screenshot of the Lottery Game

Figure S16: Screenshot of the Trust Game—Sending Situation

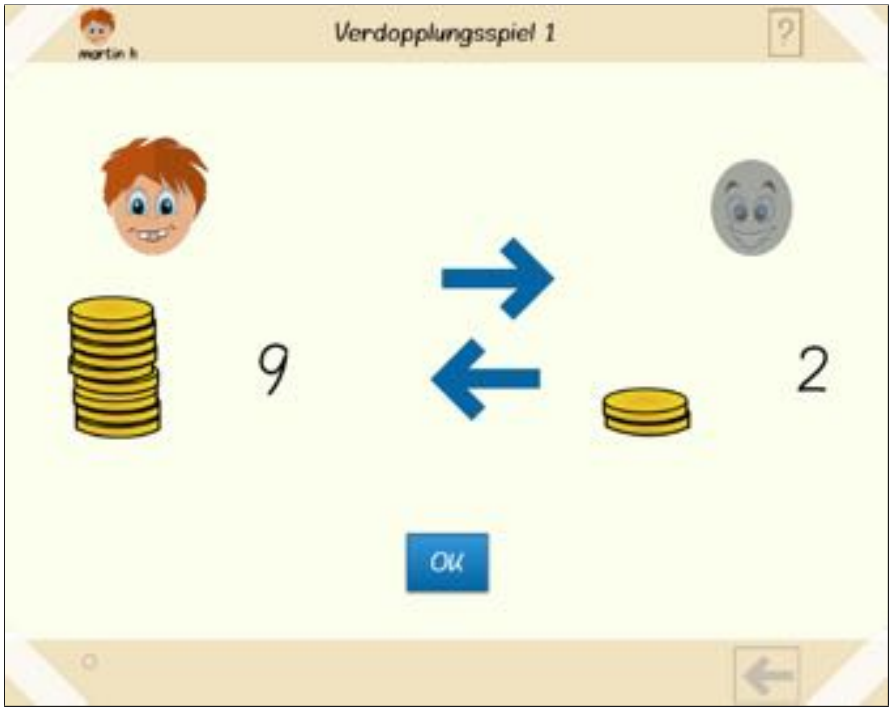


Figure S17: Screenshot of the Trust Game—Receiving Situation

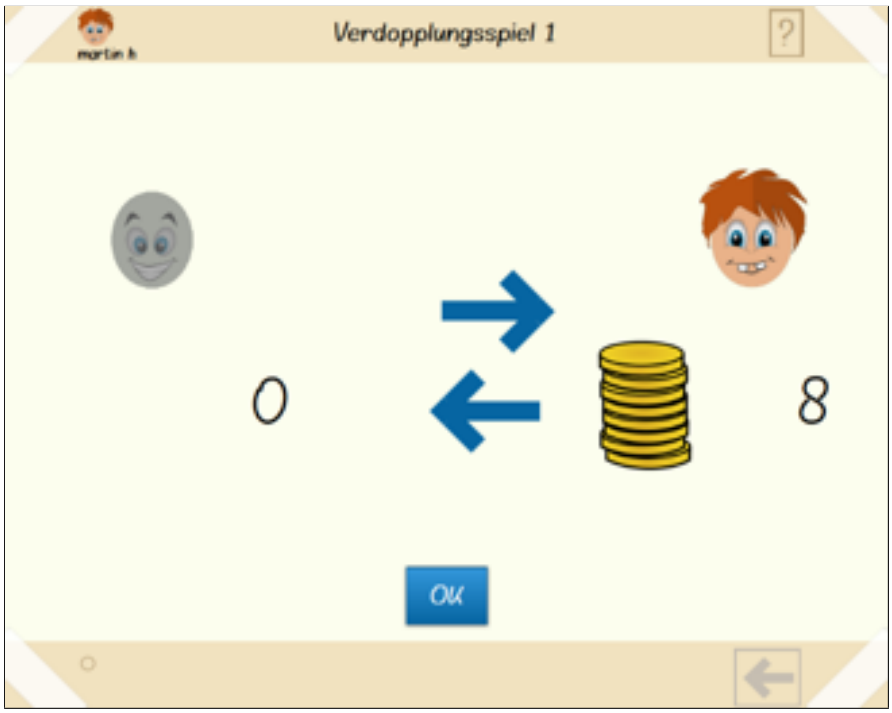


Figure S18: Screenshot of the Dictator Game

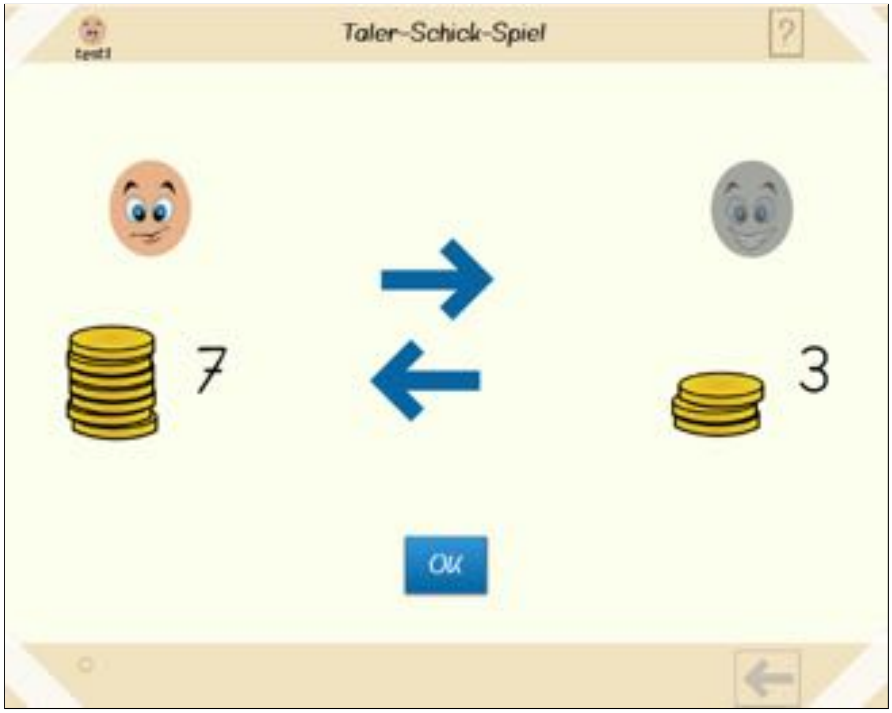
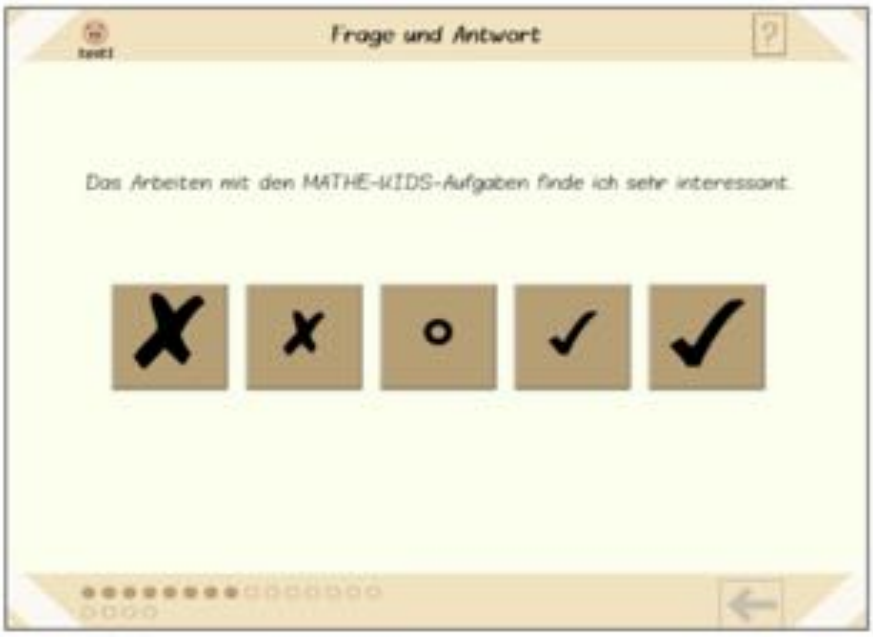


Figure S19: Screenshot of the Child Questionnaire



C Supplementary Tables

Table S1: Child Questionnaire

ID	Question
	Motivation, (Wilde, Bätz, Kovaleva, & Urhahne, 2009)
1	Math lessons are fun.
2	I think the math lessons are interesting.
3	Math lessons are enjoyable.
	Perceived Stress, (Wilde et al., 2009, pressure/tension dimension)—*only asked post-treatment*
4	I felt under pressure while doing the MATHE-KIDS tasks.
5	While doing the MATHE-KIDS tasks I felt nervous and strained.
6	While doing the MATHE-KIDS tasks I worried if I could do a good job.
7	Concerning the *Sprints* I was afraid of performing worse than my classmates.
	Somatic Problems, (Snoeren & Hoefnagels, 2014)
8	Last week I had problems sleeping (e.g. did not sleep well, woke up at night, could not fall asleep).
9	Last week I had a stomach ache.
10	Last week I was not hungry although I did not eat a lot.
11	Last week I had headaches.
12	Last week I felt tired and weak.
	Self-Efficacy, (Schwarzer & Jerusalem, 1999)
13	I am able to solve difficult problems in class. I just have to try hard enough.
14	It is easy for me to understand new topics in class.
15	Sometimes I get a bad grade. However, I know that I can become as good in school as I want to.
16	If my teacher explains topics very quickly, I cannot keep up (Reverse Scored).
	Attitude toward Competition, (Smither & Houston, 1992)
17	I like to compare myself and my performance to my classmates.
18	In games or competitions I just try not to be last.
19	In games or competitions I always try to end up first.
20	I love to compete with others.

Table S2: Teacher Questionnaire

ID	Please rate *child X* on a scale from 1 to 4.
	Motivation
1	*X* is motivated.
2	*X* has fun during lesson.
	Math Abilities
3	*X* is good at math.
	Risk
4	*X* is risk-seeking.
	Social Behavior
5	*X* behaves in a social manner.
6	*X* is cooperative.
	Grades: *Please indicate the child's grade from 1 ("very good") to 6 ("insufficient").*
5	Grade in Math for *X*.
6	Grade in German for *X*.

Notes: 1 = strongly disagree, 2 = disagree, 3 = agree, 4 = strongly agree

Table S3: Descriptive Statistics for Full Sample

	Mean	SD	Min	Max	N
Treatment	0.49	0.50	0	1	399
School 1	0.14	0.34	0	1	399
School 2	0.19	0.39	0	1	399
School 3	0.16	0.37	0	1	399
School 4	0.10	0.30	0	1	399
School 5	0.14	0.35	0	1	399
School 6	0.11	0.32	0	1	399
School 7	0.16	0.37	0	1	399
Male	0.53	0.50	0	1	399
Age (in years)	8.61	0.48	6.7	11	399
Grade Math (at baseline)	2.40	1.16	1	6	393
Grade German (at baseline)	2.60	1.14	1	6	387
Low-achiever	0.26	0.44	0	1	378
High-achiever	0.30	0.46	0	1	378

Notes: The number of observations for teacher grades is lower due to missing information on teacher questionnaires.

Table S4: Randomization for Sociodemographic Variables

	(1) Male	(2) Age	(3) Grade Math	(4) Grade German	(5) Low-Achiev.	(6) High-Achiev.
Treatment	-0.004	0.068	0.179	0.130	0.090	0.052
	(0.037)	(0.060)	(0.169)	(0.238)	(0.077)	(0.080)
School FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	399	399	393	387	378	378

Notes: OLS regression with sociodemographic variables as regressand. Standard errors in parentheses clustered on school level. We further control for school fixed effects (School FE).
* $p<.10$, ** $p<.05$, *** $p<.01$

Table S5: Randomization for Outcomes Prior to Treatment I

	(1) Motiv. Test	(2) Teac-r Motiv.	(3) Child-r Motiv.	(4) Math Add/Sub	(5) Math Mult/Div
Treatment	-0.081	-0.037	-0.029	-0.227	-0.294**
	(0.140)	(0.150)	(0.084)	(0.153)	(0.110)
School FE	Yes	Yes	Yes	Yes	Yes
Observations	378	378	376	378	378

Notes: OLS regression with baseline score of the respective outcomes as regressand. Standard errors in parentheses clustered on school level. We further control for school fixed effects (School FE).
* $p<.10$, ** $p<.05$, *** $p<.01$

Table S6: Randomization for Outcomes Prior to Treatment II

	(1) Child-r Somatic Prob	(2) Child-r Self-eff.	(3) Child-r Liking Comp.
Treatment	-0.070	-0.032	0.036
	(0.070)	(0.094)	(0.103)
School FE	Yes	Yes	Yes
Observations	375	375	375

Notes: OLS regression with baseline score of the respective outcome as regressand. Standard errors in parentheses clustered on school level. We further control for school fixed effects (School FE). * $p<.10$, ** $p<.05$, *** $p<.01$

Table S7: Mechanism—Results for Children with High Number of Points for Improvements

	Motiv. Task	Teac-r Mot.	Child-r Mot.	Vol. Tasks	Vol. Time	Math Add/Sub	Math Mult/Div
Treat × Low	0.666**	0.458***	0.096	0.570***	0.630**	0.400	0.977**
	(0.285)	(0.153)	(0.334)	(0.212)	(0.286)	(0.475)	(0.381)
Treat (Mid)	-0.239	-0.015	0.112	-0.161	-0.025	-0.206	-0.307
	(0.209)	(0.236)	(0.253)	(0.268)	(0.293)	(0.294)	(0.265)
Treat × High	0.179	0.058	0.151	-0.249	-0.322	0.057	0.252
	(0.397)	(0.207)	(0.371)	(0.312)	(0.342)	(0.336)	(0.312)
School FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Treat +	0.427**	0.442*	0.208	0.410**	0.606	0.195	0.670**
Treat × Low	(0.190)	(0.253)	(0.208)	(0.200)	(0.394)	(0.294)	(0.300)
N	183	183	177	183	183	183	183

Notes: OLS regression with post-treatment level of the respective outcome as regressand. The sample is restricted to children earning points above the class-median for days with points earned only for performance improvements. "Treat + Treat + Treat × Low" refers to the linear combination of the coefficients for "Treatment" and the interaction of "Treatment" with "Low-Achievers"; it indicates the difference between low-achieving children in the treatment group and low-achieving children in the control group. Standard errors in parentheses are clustered on class level and corrected for small number of clusters using biased-reduced linearization (BRL, Bell & McCaffrey 2002). * p<.10, ** p<.05, *** p<.01

Table S8: Mechanism—Results for Children who Improved Rank

	Motiv. Task	Teac-r Mot.	Child-r Mot.	Vol. Tasks	Vol. Time	Math Add/Sub	Math Mult/Div
Treat × Low	1.119***	0.542***	0.299	0.595***	0.574*	0.944*	1.274***
	(0.318)	(0.192)	(0.326)	(0.216)	(0.312)	(0.555)	(0.442)
Treat (Mid)	-0.384	0.071	-0.153	-0.112	0.070	-0.456	-0.569*
	(0.244)	(0.246)	(0.212)	(0.214)	(0.270)	(0.401)	(0.340)
Treat × High	0.683**	0.035	0.626*	-0.496*	-0.483*	0.467	0.552*
	(0.324)	(0.234)	(0.317)	(0.294)	(0.245)	(0.421)	(0.309)
School FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Treat +	0.735***	0.613**	0.147	0.483**	0.644	0.487	0.705**
Treat × Low	(0.199)	(0.303)	(0.274)	(0.227)	(0.424)	(0.334)	(0.299)
N	163	163	158	163	163	163	163

Notes: OLS regression with post-treatment level of the respective outcome as regressand. The sample is restricted to children who improved their average rank from days 1–7 compared with days 8–15, see Figure S1). "Treat + Treat + Treat × Low" refers to the linear combination of the coefficients for "Treatment" and the interaction of "Treatment" with "Low-Achievers"; it indicates the difference between low-achieving children in the treatment group and low-achieving children in the control group. Standard errors in parentheses are clustered on class level and corrected for small number of clusters using biased-reduced linearization (BRL, Bell & McCaffrey 2002). * p<.10, ** p<.05, *** p<.01

Table S9: Main Results for Motivation, Effort, and Math Perf. without Additional Controls

	Motiv. Task	Teac-r Mot.	Child-r Mot.	Vol. Tasks	Vol. Time	Math Add/Sub	Math Mult/Div
Treat × Low	0.394*	0.418***	0.220	0.347*	0.354	0.280	0.651**
	(0.213)	(0.155)	(0.261)	(0.191)	(0.224)	(0.291)	(0.281)
Treat (Mid)	-0.127	0.045	-0.065	-0.179	-0.063	-0.231	-0.295
	(0.152)	(0.230)	(0.213)	(0.183)	(0.187)	(0.200)	(0.206)
Treat × High	0.253	0.035	0.224	-0.213	-0.150	0.368	0.415*
	(0.206)	(0.221)	(0.251)	(0.232)	(0.212)	(0.261)	(0.248)
School FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Treat +	0.267	0.464*	0.155	0.168	0.290	0.049	0.357*
Treat × Low	(0.168)	(0.277)	(0.209)	(0.211)	(0.307)	(0.198)	(0.214)
N	378	378	361	378	378	378	378

Notes: OLS regression with post-treatment level of the respective outcome as regressand. We only control for school fixed effects (School FE) and baseline measure for outcome score. "Treat + Treat + Treat × Low" refers to the linear combination of the coefficients for "Treatment" and the interaction of "Treatment" with "Low-Achievers"; it indicates the difference between low-achieving children in the treatment group and low-achieving children in the control group. Standard errors in parentheses are clustered on class level and corrected for small number of clusters using biased-reduced linearization (BRL, Bell & McCaffrey 2002). * p<.10, ** p<.05, *** p<.01

Table S10: Main Results for Child-rated Outcomes without Additional Controls

	Perc. Stress	Somatic Probl.	Self-efficacy	Liking Competit.
Treat × Low	0.333*	-0.076	0.439*	0.599***
	(0.197)	(0.244)	(0.246)	(0.194)
Treat (Mid)	0.071	0.032	-0.115	-0.261**
	(0.191)	(0.188)	(0.136)	(0.107)
Treat × High	-0.201	-0.026	0.104	0.102
	(0.298)	(0.190)	(0.173)	(0.169)
School FEs	Yes	Yes	Yes	Yes
Treat +	0.404***	-0.043	0.324	0.338**
Treat × Low	(0.131)	(0.140)	(0.208)	(0.169)
N	363	360	360	360

Notes: OLS regression with post-treatment level of the respective outcome as regressand. We only control for school fixed effects (School FE) and baseline measure for outcome score. "Treat + Treat + Treat × Low" refers to the linear combination of the coefficients for "Treatment" and the interaction of "Treatment" with "Low-Achievers"; it indicates the difference between low-achieving children in the treatment group and low-achieving children in the control group. Standard errors in parentheses are clustered on class level and corrected for small number of clusters using biased-reduced linearization (BRL, Bell & McCaffrey 2002). * p<.10, ** p<.05, *** p<.01

Table S11: Results for Risk and Social Outcomes

	Index Risk-taking	Index Social Behavior
Treat × Low	-0.360	0.008
	(0.250)	(0.222)
Treat (Mid)	0.057	-0.106
	(0.146)	(0.177)
Treat × High	-0.147	0.086
	(0.207)	(0.188)
School FEs	Yes	Yes
Controls	Yes	Yes
Treat +	-0.303	-0.099
Treat × Low	(0.199)	(0.165)
N	370	368

Notes: OLS regressions with post-treatment level of the respective outcome as regressand. "Treat + Treat + Treat × Low" refers to the linear combination of the coefficients for "Treatment" and the interaction of "Treatment" with "Low-Achievers"; it indicates the difference between low-achieving children in the treatment group and low-achieving children in the control group. Standard errors in parentheses are clustered on class level and corrected for small number of clusters using biased-reduced linearization (BRL, Bell & McCaffrey 2002). * $p<.10$, ** $p<.05$, *** $p<.01$

Table S12: Results for Teacher Ratings

	Teac-r Math Ability	Teac-r Math Grade	Teac-r Risk	Teac-r Social Beh.	Teac-r Stress
Treat × Low	0.065	0.228	0.091	0.013	-0.050
	(0.141)	(0.174)	(0.237)	(0.225)	(0.253)
Treat (Mid)	0.168	-0.026	-0.219	0.078	0.175
	(0.134)	(0.093)	(0.171)	(0.188)	(0.236)
Treat × High	-0.038	0.042	-0.094	-0.047	-0.058
	(0.142)	(0.115)	(0.209)	(0.162)	(0.196)
School FEs	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes
Treat +	0.233**	0.202	-0.128	0.091	0.125
Treat × Low	(0.099)	(0.146)	(0.256)	(0.237)	(0.242)
N	360	378	378	378	378

Notes: OLS regression with post-treatment level of the respective outcome as regressand. "Treat + Treat + Treat × Low" refers to the linear combination of the coefficients for "Treatment" and the interaction of "Treatment" with "Low-Achievers"; it indicates the difference between low-achieving children in the treatment group and low-achieving children in the control group. Standard errors in parentheses are clustered on class level and corrected for small number of clusters using biased-reduced linearization (BRL, Bell & McCaffrey 2002).
* $p<.10$, ** $p<.05$, *** $p<.01$

Table S13: Robustness Check with Room for Improvement in Math Tasks

	Math Add/Sub	Math Mult/Div
Treatment	-0.037	-0.058
	(0.081)	(0.082)
Treat x Room	0.075	0.384
	(0.326)	(0.334)
School FEs	Yes	Yes
Controls	Yes	Yes
Treat +	0.038	0.326
Treat × Room	(0.310)	(0.341)
N	378	378

Notes: OLS regressions with post-treatment level of the respective outcome as regressand. The dummy for "Room for Improvement" is one if a child does not score 10 points in the math task at baseline. "Treat + Treat x Room" refers to the linear combination of the coefficients for "Treatment" and the interaction of "Treatment" with "Room for Improvement"; it indicates the difference between children with room for improvement in the treatment group and those in the control group. Standard errors in parentheses are clustered on class level and corrected for small number of clusters using biased-reduced linearization (BRL, Bell & McCaffrey 2002).
* $p<.10$, ** $p<.05$, *** $p<.01$