University of Zurich[UZH]

URPP Equality of Opportunity

Teaching Self-Regulation

Eva M. Berger
Ernst Fehr
Henning Hermes
Daniel Schunk
Kirsten Winkel

Equality of Opportunity Research Series #4
August 2022

URPP Equality of Opportunity

URPP Equality of Opportunity Discussion Paper Series No.4, August 2022

Teaching Self-Regulation

Eva M. Berger
German Council of Economic Experts
eva.berger@svr-wirtschaft.de

Ernst Fehr
University of Zurich
ernst.fehr@econ.uzh.ch

Henning Hermes
University of Dusseldorf
hermes@dice.hhu.de

Daniel Schunk
University of Mainz
daniel.schunk@uni-mainz.de

Kirsten Winkel
University of Mainz
kirsten.winkel@htwsaar.de

URPP Equality of Opportunity, University of Zurich, Schoenberggasse 1, 8001 Zurich, Switzerland
info@equality.uzh.ch, www.urpp-equality.uzh.ch

Teaching Self-Regulation

Authors: Daniel Schunk[*1†], Eva M. Berger[2†], Henning Hermes[3†], Kirsten Winkel[1†], Ernst Fehr[4]

Affiliations:

[1]Johannes Gutenberg University of Mainz, Department of Law and Economics, Jakob-Welder-Weg 4, 55128 Mainz, Germany.

[2]German Council of Economic Experts, c/o Federal Statistical Office, Gustav-Stresemann-Ring 11, 65189 Wiesbaden, Germany.

[3]Heinrich Heine University of Düsseldorf, DICE (Düsseldorf Institute for Competition Economics), Universitätsstr. 1, 40225 Düsseldorf, Germany.

[4]University of Zurich, Department of Economics, Blümlisalpstrasse 10, 8006 Zurich, Switzerland.

[†]These authors contributed equally to this work.

***Corresponding author**. E-mail: daniel.schunk@uni-mainz.de

Abstract:

Children's self-regulation abilities are key predictors of educational success and other life outcomes such as income and health. However, self-regulation is not a school subject, and knowledge about how to generate lasting improvements in self-regulation *and* academic achievements with easily scalable, low-cost interventions is still limited. Here, we report the results of a randomized controlled field study which integrates a short self-regulation teaching unit based on the concept of mental contrasting with implementation intentions into the school curriculum of first graders. We demonstrate that the treatment increases children's skills in terms of impulse control and self-regulation while also generating lasting improvements in academic skills like reading and monitoring careless mistakes. Moreover, it has a substantial effect on children's long-term school career by increasing the likelihood of enrolling in an advanced secondary school track three years later. Thus, self-regulation teaching can be integrated into the regular school curriculum at low cost, is easily scalable and can substantially improve important abilities and children's educational career path.

Introduction

Self-regulation refers to the capacity to regulate attention, emotion, impulses, and behavior directed at pursuing individually valued goals[1-5]. Children's early self-regulation plays a major role in their school readiness, their school achievement, and in a range of later life outcomes, such as educational attainment, income and wealth, health, and criminal behavior[6-13]. In addition, the proliferation of recently developed distance learning methods greatly increases the demands on children's self-regulation capabilities[14,15] – prompting leading institutions such as the UNESCO to conclude that self-regulation is the key 21st century skill for student success and should thus be primarily promoted[16]. A considerable literature emphasizes that self-regulation and related skills are malleable in children[17-23], especially by use of explicit strategy instruction[24-26], and that influences during early childhood and in younger ages generally play an important role in later skill formation[27-31].

Despite its fundamental importance, self-regulation is not a school subject—such as reading, language, or math—that is directly taught in schools as a regular part of the curriculum. It is at best an implicit part of normal school education that typically lacks a sufficient curricular basis. A key challenge for introducing self-regulation into the curriculum is due to the many other competing demands on scarce teaching time; adding further subjects or replacing valuable teaching time foreseen for other important topics thus entails various direct and indirect costs. Imagine, however, an effective method of teaching self-regulation that does not compete with other uses of teaching time, but which substantially improves learning in other school subjects and thus positively affects children's educational career path. Moreover, suppose that teachers could learn this method in little time by equipping them with appropriate knowledge and materials. Such a method would not only greatly reduce the opportunity cost of improving self-regulation, but would, in addition, enhance the teaching of other school subjects and be easily scalable.

Here, we propose that a short self-regulation teaching unit consisting of five lessons developed on the basis of "Mental Contrasting with Implementation Intentions (MCII)" can fit that bill[32-34]. MCII is a metacognitive strategy that addresses goal setting and goal striving as well as overcoming obstacles that prevent individuals from reaching their goals. The general idea underlying MCII is that once individuals set a goal, they imagine the positive consequences of achieving the goal, which enhances goal commitment, but they also contrast the goal with the obstacles that are currently in their way. This constitutes the mental contrasting (MC) part of MCII. Subsequently, MCII requires the identification of concrete behaviors for overcoming the barriers and of forming implementation intentions in the form of "when-then" plans that indicate a concrete self-regulatory action whenever the identified obstacle emerges. This latter part of the strategy—the implementation intention (II) part—is intended to automatize the implementation of behaviors that help overcome the obstacles.

Mental contrasting with implementation intentions provides a general method for helping individuals achieve desirable goals[32-34]. It has the advantage that it can be applied to a wide range of different goals[35]. Because these goals can also be directly related to various school subjects, MCII can, in principle, be used to enhance learning in these fields. A potential disadvantage is that it is very hard – and in the opinion of some teachers impossible – to teach the abstract concept of MCII to young children such as first graders. In fact, when we first discussed this concept with the schoolteachers, they were extremely skeptical whether MCII could be applied to first graders because children at that age have very limited abilities to understand general, abstract ideas and their reading and writing abilities are also very limited (e.g., they typically do not know all letters of the alphabet yet). In addition, children at that age are often characterized by limited goal setting skills, patience, attention span and inhibition skills, as well as a lack of perseverance, and they do not feel

responsible for their own learning progress. These limits also constrain them in transferring the method to other tasks and contexts. However, these limitations are at the same time exactly the reason why young children like first graders would particularly benefit from effective self-regulation strategies.

According to a recent meta study[36], there are no studies where self-regulation based on MCII is taught by the schoolteachers and integrated into the regular school curriculum. Two previous studies recruited sixth- and seventh graders[37] and fifth-graders[38], but MCII was provided by trained experts outside of regular classroom teaching in both studies. The first study found that parent-rated self-regulation is higher in treated children two weeks after the intervention. The second study reports that the MCII-treated children have better report card and behavioral grades at the end of the 3rd quarter—during which MCII was implemented—but these effects became small and insignificant in the 4th quarter. Thus, the longer-run effects of MCII on children's academic outcomes and school careers are basically unknown[39] and it remains unclear whether MCII can be successfully integrated into the early school curriculum to enhance not only young children's self-regulation skills but also their skills in traditional school subjects such as reading. Here, we develop a relatively brief and scalable self-regulation teaching unit based on MCII, delivered by *teachers* and integrated into the *regular school curriculum,* and we test whether it can yield sustained benefits in academic outcomes for *children in primary school*.

To do so, we conducted a randomized field experiment with 572 schoolchildren in 31 first grade classes in 12 schools in Germany (Supplementary Information, Section 1.1., Figure S1, and Table S1). In the treatment condition, the children were taught five self-regulation lessons on the basis of MCII. These lessons were spread over five weeks and directly tied to the teaching of skills that are fundamentally important for first graders—practicing reading and monitoring own mistakes. The fact that we did not apply MCII to math enables us to examine whether the taught self-regulation skills automatically extend to and improve academic skills in other—untrained—fields as well. The children's regular schoolteachers conducted the self-regulation teaching. It was embedded in everyday classroom activities and introduced to the children as part of their regular curriculum. Therefore, the children perceived it as a natural part of classroom teaching, which makes Hawthorne effects unlikely to occur.

The control group received regular classroom teaching which consisted of language lessons (reading and writing) and math lessons. Therefore, we can address the question of whether the self-regulation teaching lessons actually yield larger or smaller benefits than using scarce teaching time for the standard curriculum—a question of utmost importance for (educational) policy.

Because entire (treatment) classes are taught MCII and the control group classes continue with the standard curriculum, we randomized at the class level within schools with at least one treatment and one control class per school. This has the advantage that we can control for school fixed effects and that potential within-class peer effects of the self-regulation intervention can play a role. Consider, for example, children who often disturb in class and disrupt their peers. Self-regulation teaching could help improve these children's behavioral control and thus improve their educational performance. In addition, other children in the classroom might also benefit from a quieter classroom environment and thus also improve their educational performance. In essence, our setting allows us to evaluate the total effect of teaching self-regulation in school, including reinforcing peer effects.

In view of the challenges involved in teaching MCII to first graders, we developed five completely scripted school lessons (lasting 50 min. each) and a detailed set of materials to address these challenges (Methodology, Section A). We also instructed the teachers in a three-hour workshop

how to implement self-regulation teaching in the classroom (Supplementary Information, Section 1.2). Importantly, while teachers were instructed how to conduct the lessons, they were not informed about any specific hypotheses related to the intervention.

To assess the intervention effects, we administered standardized computer-based tests of children's self-regulation abilities as well as their academic abilities in reading and math (Methodology, Section B). The staff that conducted these tests was blind to treatment conditions, and the teachers were neither involved in the tests nor informed about their content nor the test results. In addition, we complemented these tests with teachers' assessments of the children's reading and self-regulation skills. The combination of objective, computer-based tests with teachers' ratings also enables us to check the credibility and validity of the teachers' ratings. To learn about the dynamic effect of the intervention, the outcome evaluations were carried out in four waves extending over the course of more than one year: prior to treatment (t_0), 4–5 weeks after the treatment (t_1), as well as 6 months (t_2) and 12–13 months (t_3) after treatment. All objective tests were adapted to the children's age. Furthermore, in a three-year follow-up we collected information about the children's secondary school track enrollment. The choice of secondary school track is a high-stakes educational decision in Germany, as it strongly predicts the likelihood of later enrollment at a university/college. It is therefore of direct relevance to adult labor market outcomes. If self-regulation teaching improves key skills, the trained children may have a higher propensity to move into an advanced school track (college preparatory, referred to as *Gymnasium* in German) in secondary school.

What should we expect regarding the effects of the self-regulation teaching unit? Effects may not occur directly after the teaching unit because it takes time for the children to internalize the strategy, to apply it repeatedly to different contexts, to learn from the feedback that they receive, and to get more proficient in using it. Therefore, we conjectured that the outcome measures in t_1 (assessed 4–5 weeks after the teaching unit) may not yet show clearly visible treatment effects.

In terms of outcome categories, we expected that if MCII teaching generates treatment effects, these effects are more likely to show up in domains to which MCII has been directly applied—reading skills and the ability to find careless mistakes (outcome category 1). We also conjectured that it might enable the children to better inhibit prepotent impulses and improve their self-regulatory classroom behaviors (outcome category 2). This conjecture follows from the fact that MCII represents a self-regulation strategy that requires the children to approach goal implementation in a systematic manner by overcoming obstacles that often come in the form of strong temptations. In contrast, we were considerably more pessimistic about the children's ability to automatically generalize and extend the strategy to other academic subjects or other domains. It is, perhaps, too much to expect first-graders to already have the cognitive capacity for abstract thinking and generalization that these automatic extensions to other academic domains require.

With regard to the impact of MCII teaching on children's longer run school career path, we remained entirely agnostic. In this context, it is important to keep in mind that previous studies on MCII teaching in school children only reported very short run effects[37] or effects that vanished in the next school quarter[38]. Thus, showing a sustained effect of the teaching unit after 6 months (t_2) and after 12-13 months (t_3) goes already considerably beyond the previously available evidence. While it is definitely possible that a short run intervention like ours triggers a process that benefits the children several years after the teaching unit, it is also entirely possible that the benefits deteriorate and vanish.

Results

Randomization and Sample Balance

The randomization into treatment and control group led to a balanced sample, as documented in tests for differences between treatment and control group conducted by regressing various socio-demographic background variables measured at baseline (t_0) on the treatment dummy (Supplementary Information, Table S2). Similarly, we test for imbalances in our outcome measures prior to treatment (Supplementary Information, Table S3). Overall, there is no evidence for imbalances between treatment and control group beyond differences caused by chance; moreover, we control for any residual nonsignificant imbalances in our econometric analyses by controlling for the children's baseline characteristics (Supplementary Information, Section 1.5, for more details).

Main Results

The following results are based on OLS regressions that regress the respective outcomes (e.g., reading abilities displayed in the reading test) at three different points in time—at t_1 (4–5 weeks after the treatment), t_2 (6 months after the treatment), and t_3 (12–13 months after the treatment) —on a treatment dummy and control variables. As we stratified our randomization on the school level, we include school-fixed effects. Doing so removes noise that is due to school facilities or social background differences between schools. To increase the precision of the estimated treatment effect, we also include the respective baseline outcome score (measured prior to treatment at t_0) in each regression as a control variable (Supplementary Information, Section 1.5). It has been shown that this method provides more precise results than the difference-in-differences estimators that compare the outcome *changes* from pre- to post-teaching measures between treatment and control groups[40,41]. We allow for interdependence of observations within classrooms by clustering the standard errors at the classroom level, and we also report p-values that are adjusted for multiple hypothesis testing[42] (Supplementary Information, Section 1.5). All outcomes are standardized in order to make treatment effects comparable in size.

MCII teaching already has a significant effect in t_1 on the reading test (0.20 SD, $p = 0.020$, Fig. 1, Table S4), but this effect is somewhat fragile as indicated by the larger confidence intervals in t_2 (0.21 SD, $p = 0.111$). However, the treatment effect in t_3 becomes sizeable and highly significant (0.39 SD, $p = 0.006$). Although the teachers were blind to all computer-based tests, a similar picture emerges from the teachers' assessment of the children's overall reading abilities. They indicate no treatment effect in t_1 (0.002 SD, $p = 0.983$), a treatment effect in t_2 that just passes the 5% significance threshold (0.288 SD, $p = 0.049$) and again a sizeable and robustly significant effect in t_3 (0.366 SD, $p = 0.005$). It is reassuring that the teachers' assessments of overall reading ability are quite consistent with the results from the objective reading test, even though the teachers were not involved in the reading test and did not know its results, suggesting that demand effects do not drive teachers' assessments.

Further evidence for the credibility of teachers' assessments is provided by the strong correlation (Spearman's rank correlation, $\rho = 0.78$, $p < 0.001$) between the children's average score in the four objective reading tests (in t_0, t_1, t_2, and t_3) and the teachers' average reading assessment of the children. In addition, we observe that the teachers' ratings in the first assessment after the self-regulation intervention (in t_1) are even more conservative than the results of the objective computer-based reading tests. In the presence of demand effects, one would expect the opposite result, i.e., that the teachers report overly optimistic reading assessments. The teachers' overall assessment of children's ability to find careless mistakes follows a similar time pattern as their assessment of the

overall reading ability (Fig. 1): there is no treatment effect in t_1 (0.025 SD, $p = 0.858$) but significant and increasing treatment effects in t_2 (0.474 SD, $p = 0.013$) and t_3 (0.691 SD, $p = 0.001$).

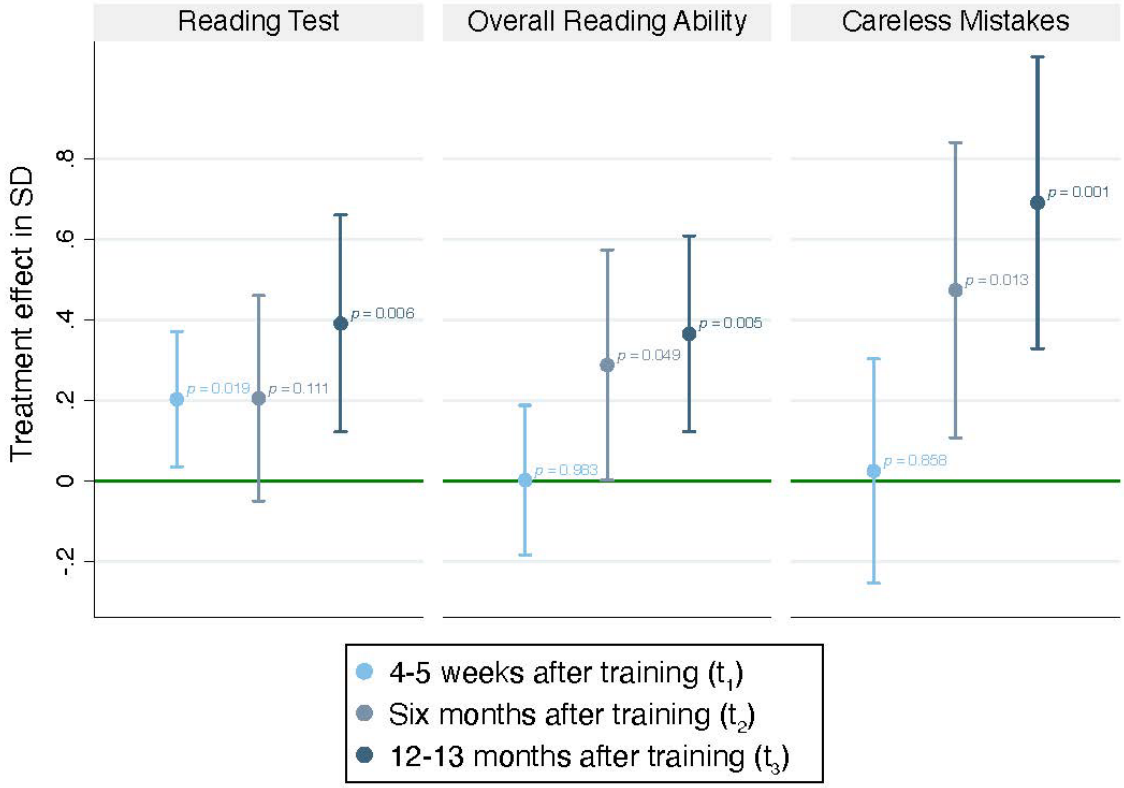


Figure 1--- The effects of self-regulation teaching on reading and finding careless mistakes. The dots show point estimates (as fractions of a standard deviation of the respective outcome) of how MCII teaching changes the outcome indicated in the panel title relative to the control group. *Reading test* is the score from the reading comprehension test, *Overall reading ability* indicates the teachers' overall assessment of the children's reading abilities. *Careless mistakes* shows the teachers' assessment of the children's ability to find careless mistakes. All p-values refer to two-sided t-tests. The bars indicate 95% confidence intervals. All estimates are based on least squares models controlling for school fixed effects, pre-treatment outcome scores, and further controls (see Supplementary Information, Sections 1.5 and Table S4, for more details and p-values adjusted for multiple hypothesis testing). Standard errors are clustered at the classroom level.

The above-reported effect sizes are quite substantial as it has been pointed out that "in real-world settings, a fifth of a standard deviation (0.2 SD) is a large effect"[43]. This assessment is supported by 'the best evidence synthesis' literature[44] which suggests the use of empirical benchmarks from high quality field research on education instead of benchmarking on the basis of laboratory studies[45,46]. A comparison of our results with the control group's scores provides another intuitive benchmark for assessing the effect size. For example, if we compare the treatment effect on the reading score in t_3 to the distribution of the control group's reading scores, we find that the effect size of 0.39 SD moves the median child's reading score in the control group from the 50th to the 75th percentile. For the careless mistakes' outcome, the treatment effect is very similar in size, moving the median control group child again from the 50th to the 75th percentile. Thus, taken together, these

results suggest that the application of five lessons of MCII teaching to reading and finding careless mistakes causes significant and sizeable outcome improvements one year later in these domains.

How does MCII teaching affect the ability to inhibit pre-potent impulses ("inhibition"), the ability to attend and quickly respond to stimuli that require an action ("attention"), and overall self-regulation ability as assessed by teachers (outcome category 2)? We find a significantly positive treatment effect (Fig. 2, Table S5) on inhibition (measured by the negatively signed commission errors in the go/no-go task; 0.26 SD, $p < 0.001$) and attention (measured by the negatively signed omission errors in this task; 0.56 SD, $p < 0.001$) 12–13 months after the treatment (t_3). Interestingly, as with the measures in outcome category 1, the effects are weaker and non-significant 4–5 weeks after the teaching unit (in t_1), suggesting that the teaching needs time to come to fruition.

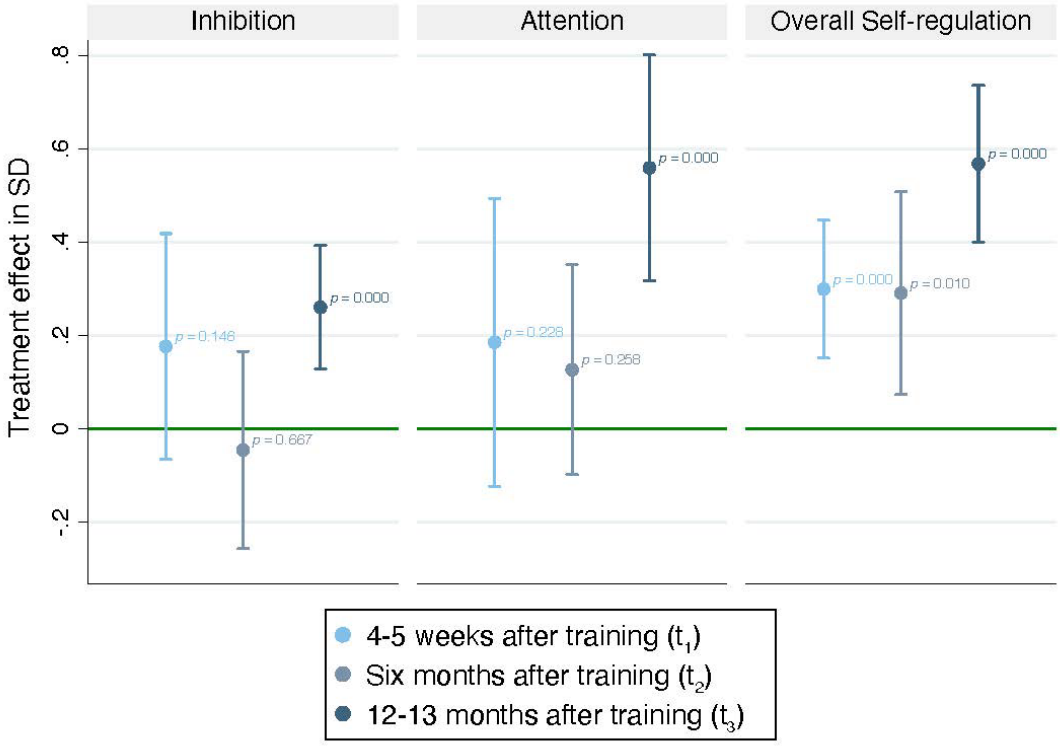


Figure 2 --- The effects of self-regulation teaching on inhibition, attention, and overall self-regulation. The dots show point estimates (as fractions of a standard deviation of the respective outcome) of how MCII teaching changes the outcome indicated in the panel title relative to the control group. *Inhibition* measures the children's ability to inhibit prepotent impulses (by the negatively signed commission errors) in the go/no-go task, *Attention* indicates the ability to respond properly to the stimuli that require an action (measured by the negatively signed omission errors) in the go/no-go task. *Overall self-regulation* indicates the children's assessment of their everyday self-regulation behaviors in the classroom by the teachers. All p-values refer to two-sided t-tests. The bars indicate 95% confidence intervals. All estimates are based on least squares models controlling for school fixed effects, pre-treatment outcome scores, and further controls (see Supplementary Information, Sections 1.5 and Table S5, for more details and p-values adjusted for multiple hypothesis testing). Standard errors are clustered at the classroom level.

The teachers' assessments of the children's overall self-regulation behavior in the classroom displays a roughly similar time pattern: the treatment effect is significant and largest after 12–13 months (0.57 SD, $p < 0.0001$) and similar in size to the effect on attention, yet the treatment effects

in the previous waves (t_1 and t_2) are already significant due to smaller standard errors (t_1: 0.30 SD, p = 0.002; t_2: 0.29 SD, p = 0.005) and slightly larger effect sizes compared to the inhibition and attention outcome. Thus, both the results from standardized computer-based tests and the findings from teachers' ratings suggest that the treatment improved the children's self-regulation abilities. In addition, we also collected parent ratings of their children's self-regulation six months after the training (in t_2). However, the parents' response rate was, unfortunately, considerably lower (63 percent) compared to the teachers' response rate (92 percent). The parent ratings also suffer from several other problems adding noise to the measurement (Supplementary Information, Section 1.4 and Table S15). Albeit these problems may prevent us from finding significant results, if the parent ratings point in the same direction as the other self-regulation measures, they complete the overall picture. We indeed find that parent-assessed self-regulation skills are 0.13 SDs higher in the treatment group but the effect is not statistically significant (p = 0.144; see also Supplementary Information, Section 1.4 and Table S15). Nevertheless, by pointing in the same direction they are consistent with the other results on self-regulation.

To what extent does MCII teaching spill over to an academic domain that was not targeted by the teaching unit or leads to an increase in stamina in a tedious letter detection task? We address these questions with two outcome measures—children's math skills (measured by arithmetic and geometry tests) and the letter discrimination task that requires stamina and frustration tolerance (Supplementary Information, Section 1.4). Here, we find that MCII teaching has basically no impact at all on these outcomes (Supplementary Information, Fig. S14 and Table S6). Moreover, there is no time trend whatsoever across all outcome evaluation waves: the treatment effect for these outcomes is always close to zero, suggesting that first graders do not automatically generalize the MCII teaching to new academic domains or to tedious tasks that require stamina and high frustration tolerance.

Does teacher quality or experience affect the treatment effect? This question is relevant, as more experienced teachers are typically better at educating children[47] and therefore might also have been better in teaching MCII. However, the fact that we developed detailed and fully scripted lessons for teaching MCII made it easy for the teachers to teach and apply MCII, and this may have mitigated effects of teacher experience on the treatment effect. Indeed, if we control for teacher experience (Supplementary Information, Table S8) we find that classes with teachers with a below-median experience do not show a significantly lower treatment effect. In addition, we also do not find heterogeneous treatment effects for demographic variables such as gender, age, and migration background. Moreover, MCII teaching benefits children with low *and* high self-regulation abilities at baseline alike. It is in this regard different from the effects of growth mindset interventions whose effects seem to occur primarily in low achieving children[48].

Effects on Secondary School Track Choice Three Years after Treatment

Given that we found treatment effects on important outcomes in a one-year follow-up, the question arises whether the MCII teaching has an even longer run effect on a high-stakes outcome. We therefore evaluate its effect on secondary school track choice three years after the MCII teaching—a very important and far-reaching educational decision. It turns out that children in the treatment group are 13.3 percentage points more likely to choose the advanced track (Fig. 3 and Table S7, column 1, p = .006) if we estimate the treatment effect with a linear probability model. The result is very similar when we estimate a probit model.

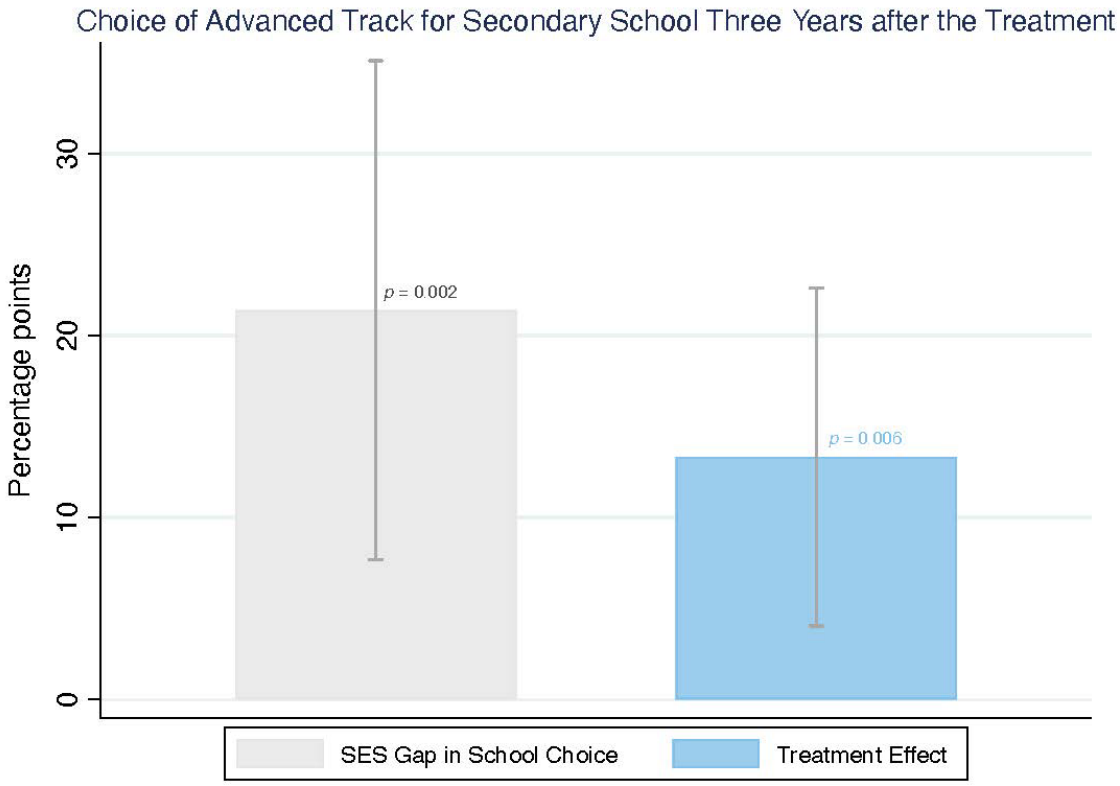


Figure 3 --- The long-term effects of self-regulation teaching on children's enrollment in advanced secondary school track. The left bar shows the socio-economic gap in enrollment to the advanced track of secondary school based on whether the child's mother has a university degree for the control group (controlling for baseline IQ). Children whose mother has a university degree are 21.4 percentage points more likely to enroll in the advanced track. The right bar shows the estimated 13.3 percentage point increase in enrollment generated by the MCII teaching (based on Table S7, Column 1). The estimate is based on a linear probability weighing model controlling for school fixed effects and further controls (see Supplementary Information, Sections 1.5 and Table S7) but the results are robust to using probit or inverse probability weighting models. All p-values refer to two-sided t-tests. The bars indicate 95% confidence intervals. Standard errors are clustered at the classroom level.

To benchmark the size of this effect, we compare it to the difference in enrollment in the advanced track by maternal education in the control group (controlling for baseline IQ): children whose mothers have a university degree are 21.4 percentage points more likely to be enrolled in an advanced track secondary school compared to those whose mothers do not have a university degree. Thus, the treatment effect of MCII teaching is roughly 13.3/21.4 = 62% as large as this important socioeconomic gap.

Because there is attrition in parents' participation in the survey that asks about school track choice, we also examined the robustness of our findings by testing whether attrition is related to treatment assignment. Fortunately, however, this is not the case: If we regress participation in the survey on the treatment condition, gender of the child, age of the child and school fixed effects we do not find significant effects. In addition, we control for attrition by inverse probability weighting and show that the treatment effect on school track choice is robust (Table S7, column 3, $p = 0.001$).

The results described above indicate that MCII teaching caused improvements in outcome category 1 (Fig. 1) and outcome category 2 (Fig. 2). Is the improvement in outcome measures in

these two categories a reason for the significantly higher enrollment of the treated children in an advanced track secondary school? To examine this question, we conducted a mediation analysis and included the t_3-values of these outcome measures as regressors for the school track choice (Table S7, column 2 and column 4). We indeed find that the children's performance in the reading test (p = 0.028), their ability to find careless mistakes (p = 0.030), and their overall self-regulation ability (p = 0.025) in t_3 are significant mediators of the treatment effect on school track choice. Reading ability as indicated by the reading test in t_3, in particular, seems to be a strong mediator of school track choice—a one SD increase in reading ability is associated with a 15.3 percentage points increase in advanced school track choice. However, a one SD improvement in finding careless mistakes or in overall self-regulation is also associated with substantial increases in enrolling in the advanced track by 6.3 and 8.3 percentage points, respectively. Moreover, the treatment variable is no longer significant if we include the children's abilities in t_3.

Discussion

Self-regulation is generally thought to be of fundamental importance for children's educational and lifetime success. There is also reason to believe that the earlier schoolchildren acquire self-regulation skills, the more they benefit from them in the long run. However, how can the teaching of effective self-regulation to young schoolchildren be integrated into their school curriculum without reducing other productive uses of teaching time? Is it possible to teach self-regulation in a way that it even substantially improves children's performance in core school subjects and thus has the potential to affect their educational career path? And how can this teaching method be designed to render it easily scalable to a larger subject population? As an answer to these questions, we have proposed a few self-regulation teaching lessons that are based on MCII.

We conducted a randomized-controlled field experiment involving 572 first graders that overcomes the challenges of teaching MCII-based self-regulation to first graders. The findings indicate that five self-regulation teaching lessons spread over five weeks can be used to generate substantial improvements in academic skills—such as reading—that are part of the standard curriculum. In addition, we show that teaching self-regulation has far transfer effects on general inhibitory and attentional abilities and improves the children's overall self-regulation behavior in the classroom.

We do not observe fade-out effects for the positively affected skills. Potential reasons for the observed sustainability are that the skills we address are thought to be not only malleable but also fundamental[49] in the sense that they are crucial for the further development of self-regulation (self-productivity) and they increase the productivity of other skill investments (dynamic complementarity)[50]. Moreover, our intervention differs in important aspects from the two MCII studies in a school context mentioned above[37,38]: our intervention is more intensive (five hours), it is conveyed in a playful, vivid, and meaningful manner, and we apply it not only to one but to several different goals, making it more likely that children will internalize the meta-cognitive strategy, thus enhancing self-regulation behavior at school in general. By addressing basic literacy skills and the monitoring of careless mistakes in particular, we also directly target skills that are fundamental for subsequent learning progress, both within and beyond the domain of reading. A distinguishing feature of our intervention is also that we randomized between (and not within) school classes. Hence, we take advantage of beneficial peer or classroom effects that may lead to a subsequently enriched environment in the treated classes, which may be crucial for sustaining earlier skill gains[49]. Positive peer group effects appear particularly plausible in view of the fact that the children stay together in the same class for four years in primary school. Overall, this sustainability translates into

a striking effect on children's school career choices three years after the MCII teaching—making it considerably more likely that they will be enrolled in an advanced track secondary school, which is known to deeply affect the children's life-time education and labor market trajectory.

Despite all our efforts to provide reliable and robust evidence, we acknowledge certain limitations of our study. First, some of our outcome measures were rated by teachers who were not blind to treatment condition. In this context it is important to emphasize that our main conclusions are based on both standardized computer-based tests of, e.g., reading ability and teachers' ratings of children's reading ability. Moreover, the strong correlation between the results of the objective tests and the teachers' ratings makes us confident about the reliability of our measures. Also, the treatment effects measured using objective tests and using teachers' ratings are very similar with regard to effect size and temporal patterns. Nevertheless, while these patterns provide little reason to doubt the validity of the teachers' ratings, we cannot fully rule out that they may contain some bias.

Second, the sample from which we draw inferences adds limitations. Especially in light of the classroom-level randomization, our sample size is limiting, e.g., the analysis of heterogeneous treatment effects. We also target a specific (and challenging) age group (first graders) in a specific education system of a developed country (Germany). Further research is necessary to learn whether the findings also hold in different age groups, in other education systems, or for settings in developing countries.

Finally, while the number of lessons replaced by our self-regulation teaching unit is very small and therefore, in our view, negligible, we do not have perfect control over the amount of time that treatment vs. control classes spent practicing reading or learning to find careless mistakes. We deem it highly unlikely that a few additional lessons of practicing reading would yield these large and long-term effects and prefer the interpretation that children learned a self-regulation strategy that helped them to improve their learning and goal striving over the following three years.

The implementation of the teaching lessons is associated with very little cost per child, as the teaching hours require only a few hours of training for the teachers and five teaching lessons for the children. Moreover, they yield high benefits even if we make rather conservative assumptions by only counting the benefits from improved reading abilities and neglecting improvements in overall self-regulation, inhibition control, or the finding of careless mistakes (Supplementary Information, Section 1.6). If we consider only the benefits of improved reading skills that already accrue one year after the MCII teaching, the cost-to-benefit ratio is 1:1.5, meaning that the benefits amount to €1.5 for every Euro spent. If we take a longer run perspective and calculate the increased lifetime earnings from improved reading skills, the cost-benefit ratio is even in the range of 1:10.

In addition to its very favorable cost-benefit ratio, the proposed method of teaching self-regulation is also easily scalable to a much larger population, as there is little reason to believe that the fully scripted self-regulation lessons we developed could not be applied to other first graders. All it takes is as little as three hours of training for the teachers to render them able to apply the method. Finally, the findings also indicate that—at least among first graders—self-regulation teaching did not automatically transfer to other academic subjects like math in the one-year period after the intervention. However, if it is possible to apply self-regulation lessons to the teaching of reading skills, we see little reason why it should not be possible to apply the lessons to teach foreign languages or other academic subjects. In fact, synergistic benefits might arise if MCII-based self-regulation teaching is applied to more than one academic field. Future research may thus extend the self-regulation teaching unit to other areas such as math or science. Additionally, collecting information on the detailed time use of the control classes, more "active" control conditions, as well

as detailed data on the use of the self-regulation strategy after the intervention has ended would be useful to learn about the specific mechanisms underlying the treatment effects.

Methodology

The study was conducted in primary schools in Mainz, Germany in 2013/2014. It consisted of a five-week intervention, four data collection waves, and a long-term follow-up survey three years after the intervention. Our study received ethical approval from the Human Subjects Committee of the Faculty of Economics, Business Administration and Information Technology at the University of Zurich in September 2012. We confirm that we have complied with all relevant ethical regulations.

In the context of a large school project[51], we recruited 12 schools with 31 classes for the study. There were 599 children in these classes in November 2012. We received 580 parental consent forms that allowed us to collect data in evaluation waves t0–t3, resulting in a consent rate of 96.8%. We were able to evaluate 572 children of the 580 for whom we received parental consent to collect data for our final data set. The children we could not evaluate either switched to non-participating classes or schools, moved away, or were ill for a longer period of time during data collection; we did not exclude any available data. Among the sample of 572 children, 292 were girls (51%) and 280 were boys (49%). Mean age prior to the intervention (Jan 2013) was 6.84 years (SD = 0.36 years). All children received a small toy for participating in the evaluation waves. We did not pay a financial compensation to children for their participation.

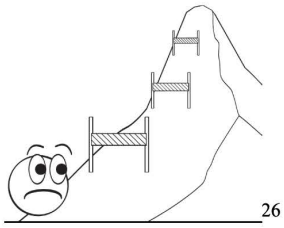
A. Addressing the challenges of teaching MCII to first graders

To address children's limited ability for abstract thinking, we developed an illustrated storybook with an appealing main character named "Hurdy", the hurdle jumper. The story unfolds in an emotionally involving way so that the children quickly identify with Hurdy, whose first goal is to climb to the top of a high mountain (Fig. 4). Hurdy imagines the great view he[52] will enjoy from the top of the mountain but contrasts this goal with the many hurdles he faces along the way. Hurdy's when-then plan is that "when he faces a hurdle, then he jumps over it". In this way the abstract MCII strategy is conveyed in a playful manner; it becomes concrete, vivid, and meaningful for the children. This enables us to use the main character's ideas and actions as a role model that helps us transferring the strategy to further goals, obstacles, and plans. [53]

(a) Introducing Hurdy, and how he imagines the great feelings he would experience when enjoying the spectacular view at the top of the mountain.

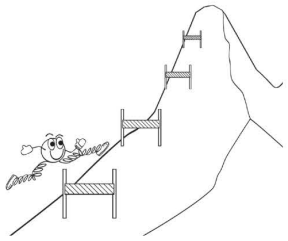
"Once upon a time, there was a small, cheerful ball. He wanted to go all the way to the top of a high mountain. He closed his eyes and imagined how wonderful he would feel when he would get aaaall, all the way to the top. He might be above the clouds, maybe even a bit closer to the stars. He imagined how wonderful the view would be up there: a panorama over broad fields, the long river, the many houses and cars that would surely look like tiny little toy cars from all the way up here. He knew: way up on top of the mountain, he would be the happiest little ball in the whole world!"

(b) Identifying the obstacles.



"But the mountain was very steep, and the ball would certainly roll down again and again. Moreover, there were lots of hurdles that he could not go over. And so the ball rolled sadly back and forth at the base of the mountain. He thought and thought again and again about how he could get to the summit of the mountain. And then! He suddenly had an idea how he could acquire legs and arms that enable him to walk, run and jump!" (*for full text see endnote 26*)

(c) Overcoming the obstacles.



"At the base of the mountain, he picked up speed, hopped a couple of times, and jumped as high as he could and—suddenly he jumped over the first hurdle. And he jumped like this further up the mountain, and hopped over the next and the next, and then all the other hurdles. It was very tiring for the little ball, but he did not give up. He was so happy that he was getting closer and closer to his goal—and he thought how beautiful it would be when he finally reached the top. He hopped further and further up the mountain and jumped over one hurdle after the other …. Until he was aaaall, all the way at the top."

(d) Enjoying goal achievement.

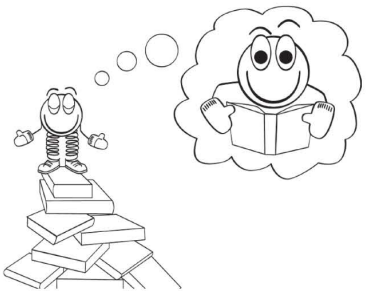
"He enjoyed the heavenly view and was so very happy that he had almost forgotten his great effort. The view was even broader and more breathtaking than he had ever imagined it could be. He lifted his arms into the air, and even though he had reached the summit, he jumped for joy into the air again and again. He was so happy and was sure that he now could jump over every hurdle that might be in his way—a wonderful feeling!"

Figure 4 --- Introducing generic components of MCII to first graders with the help of an emotionally involving story about Hurdy, the hurdle jumper. (a) Imagining a desirable goal. **(b)** Identifying obstacles and developing a solution. **(c)** When-then rule: Whenever there was a hurdle, Hurdy jumped over it. **(d)** Enjoying goal achievement. All scenes (a) – (d) are communicated with the help of a storybook containing both the pictures (on the left) and the text (on the right) that the teacher read aloud in a stepwise manner. After each step in (a) – (d), the children discussed the story in the classroom.

Once the general idea behind MCII was playfully introduced (Fig. 4), the children subsequently applied it to three goals. To practice the MCII strategy and account for the children's limited goal setting skills, the first two goals were set by us. The first goal was to become better in reading by practicing reading out loud, because reading is a skill that is fundamental for all other subjects taught in primary school. The second goal was for the children to make fewer careless mistakes in their own schoolwork by using a self-monitoring technique—the detection (and correction) of own mistakes. We used this goal because the lack of metacognitive self-monitoring strategies has been put forward as a major factor explaining cross-country differences in academic achievements in the PISA study[54]. The third goal was individually chosen by each child.

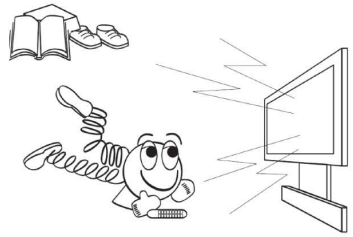
Every new goal was introduced with the help of the main character, Hurdy. For example, Fig. 5, illustrates how we used Hurdy as a role model for the application of MCII teaching to the reading goal. The teacher began by reading aloud a story where Hurdy imagines how wonderful he would be to be able to read (Fig. 5a). After the teacher had read the story, the children themselves publicly discussed what they would enjoy most if they were able to read well. Likewise, after the teacher read aloud about the obstacles that Hurdy faced (Fig. 5b) or the when-then rule that Hurdy developed (Fig. 5c), the children subsequently discussed publicly the hurdles they face themselves and possible when-then rules that help them overcome their obstacles.

(a) Imagining the joy associated with becoming a better reader.



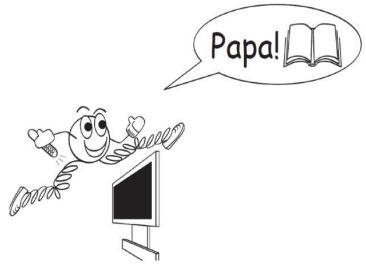
"Once Hurdy noticed how wonderful it was to reach his own goals after crossing a lot of hurdles, he soon identified his next wish. He really, really wanted to be able to read well. He wanted to be able to read all the exciting stories about the "Wild Soccer Kids" on his own. Most of all, late at night with a flashlight under his cover. He did not just want to look at the pictures in his new dinosaur book, but wanted to read the names of the dinosaurs and how they once lived on their own. And he dreamt about how proud grandpa and grandma would be of him if he could read aloud to them. Hurdy knew that he would have to practice reading a lot, but he really wanted to do it. It was his big dream."

(b) Identifying the obstacle that prevents Hurdy from becoming a better reader.



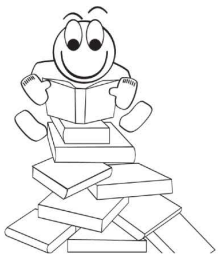
"But there were hurdles to jump over on Hurdy's journey to his new reading goals. On the way home from school, he often planned to practice a lot of reading. But then, when he got home, the big television was in the middle of the room. Hurdy took his shoes off, left his book lying next to the shoes, and turned the TV on. He had already forgotten what he had planned. The book was only lying around, and Hurdy did not worry about it anymore."

(c) Overcoming the obstacle for becoming a better reader with a when-then rule.



"When Hurdy was lying in front of the TV, he suddenly remembered how wonderful he felt when he imagined how it would be if he could read really well. How his grandma would have praised him and how he thought he would feel if he could read exciting stories under his cover at night with the flashlight. He shouted loudly, *"Papa, can I read something to you?"*, turned the TV off, ran over to his father, grabbed his book, and read something aloud to him. He realized that he could turn the TV on again after reading and watch in peace with a better feeling. He then resolved: *"**When** I want to watch TV, **then** I will first call my father, read him something aloud, and watch TV afterwards"*. In this way, Hurdy jumped over at least one TV hurdle a day and became much better in reading."

(d) Enjoying goal achievement



"Hurdy was then proud of himself because he had jumped over so many hurdles and was thus very good in reading. He read so much that he collected a real mountain of books from which he could always grab a book and start reading. The greatest pictures played inside his head, and he could see all of the stories with his inner eyes. When he read about heroes' adventures in the stories, he felt as if he had been a part of the adventure himself. His father and mother were very proud of him, and grandpa and grandma even more so."

Figure 5 --- Applying MCII to the goal of improving reading abilities (a) Imagining a desirable goal. **(b)** Identifying obstacles. **(c)** Developing and implementing a solution for overcoming the obstacle. All scenes (a) – (c) are communicated with the help of the picture on the left and the text on the right that the teacher read aloud. In addition, the teacher also read a story **(d)** about how Hurdy enjoyed the success of becoming a good reader. After the teacher had read a textbox, the story was discussed in the classroom, and the children contributed their own imaginations, obstacles, and ideas to the context. In addition, the children applied each step (a) – (d) to their situation by drawing pictures in a workbook (see Fig S1-S2) that expressed their individual imaginations, obstacles, and when-then rules

The use of Hurdy as a role model helps us transfer the MCII strategy across different goals while addressing the children's limited transfer capabilities. To further deal with this issue, we applied a scaffolding method that gradually reduces the level of support in the application of MCII. The children's obstacles and plans thus become more and more personalized from goal one to goal three, implying an increasing need for own transfer thinking. In this context, classroom discourse also played an important role because it served the purpose of fostering the transfer of the MCII components from the role model's thoughts, actions, and plans to the children's individual context. For example, after the children listened to the short story describing Hurdy's obstacle towards becoming better at reading (Fig. 5b), the subsequent classroom discourse induced the children to undertake a first small step of applying the obstacle identification component of MCII to their own situations.

To further practice and personalize the application of MCII, each child received a prepared workbook which visualized the different steps of the MCII strategy. The workbook also contained space so that the children could apply the strategy to their individual context with their own added drawings (Fig. S1 and S2). For example, children drew their ideas of the positive consequences of reaching a goal or of their individual obstacle after discussing it with classmates. The visual structure in combination with the individual drawings enables the children to internalize the MCII strategy without requiring reading or writing skills. The children thus experienced a diverse set of interesting tasks during the MCII teaching lessons—listening to Hurdy's story, discussing with their classmates, individualizing their goals in their workbooks—that kept them interested and compensated for their limited attention spans.

We addressed children's limited perseverance by spreading the five MCII teaching lessons over five weeks during which we encouraged them to pursue progressively more ambitious sub-goals related to reading and monitoring their mistakes. To constantly remind them of the different steps of the MCII strategy, a large poster that looks exactly like the first figure in their workbook (Fig. S1) remained on the wall in their classroom during the five weeks. In addition, flash cards (Fig. S3) were attached to the poster that reminded the children of the current goal, obstacles, and plan.

In principle, we could have involved the parents into the teaching and application of the MCII strategy. However, we deliberately wanted to avoid this for three reasons. First, involving the parents complicates the intervention, making it more expensive and less easily scalable. Second, if the parents take responsibility for implementing parts of MCII, the children's self-responsibility for their learning may be undermined. Because we wanted to foster their self-responsibility, the story is based on *Hurdy's desire* to reach the top of the mountain or become a good reader. Likewise, it is Hurdy who wants to become a good "error detective" (i.e., find careless mistakes), and the children's third goal was entirely self-determined. Third, involving parents might introduce heterogeneous treatment effects that depend on parents' socio-economic characteristics—a possibility that we wanted to avoid.

B. Measuring the effects of self-regulation teaching

To evaluate the effects of self-regulation teaching, we measured four types of outcomes. First, we are interested in outcomes related to the first two goals the MCII strategy was applied to—the reading goal and the goal of monitoring and correcting one's own mistakes. We measured reading comprehension skills with an objective computer-based reading test (Supplementary Information, Section 1.4) and, in addition, teachers assessed the children's *overall* reading abilities. The teachers also assessed the extent to which the children committed careless mistakes during their usual

classroom sessions. These measures allow us to answer the question whether MCII is more effective than usual classroom teaching in fostering children's abilities in domains to which MCII has been directly applied. If this was the case, MCII would be directly useful in achieving the goals of the standard curriculum.

Second, we are interested in outcomes that measure more general self-regulation skills that are not explicitly taught in the MCII teaching lessons. These are skills such as the ability to inhibit pre-potent impulses and to pay attention—measured by an objective computer-based go/no-go task (Supplementary Information, Section 1.4)—as well as an overall teacher assessment of children's self-regulation and discipline in the classroom. In the go/no-go task, the children need to attend to rapidly emerging and vanishing pictures of different animals; they have to click a button for all animals (the "go animals") except for one (the "no-go animal") within the short time period during which the animal is on the screen. Because most of the time "go animals" appear on the screen, the children are tempted to constantly push the button. However, a "no-go animal" appeared occasionally on the screen, and then they had to refrain from pushing the button. Pushing the button for "no-go animals" indicates thus a failure to inhibit a prepotent response (commission error), while not pushing the button for a "go animal" can be interpreted as an attentional failure (omission error).

Overall self-regulation in the classroom was measured with items such as "The child often disturbs class instruction" or "The child has trouble waiting until it is his/her turn" or "The child has a lot of self-discipline". The answers to these items are aggregated into an overall self-regulation index (Supplementary Information, Section 1.4). Notice that we do not train *general* inhibitory or attentional abilities like those required in the go/no-go task during the application of MCII to reading and careless mistakes. Likewise, the teaching lessons do not directly prevent children from disturbing class instruction or inducing them to be more patient until "it is his/her turn". A treatment-induced improvement in these outcomes therefore indicates far transfer effects.

Third, we want to examine whether the taught MCII strategy automatically spills over to other academic domains that self-regulation teaching did not target. This helps answer the question whether first graders automatically apply the strategy to novel academic domains. In this context, we measure whether MCII teaching improved children's math skills. In addition, we measure their stamina in a tedious and frustration-inducing letter discrimination task. In this task, the children saw a long string of different letters on the screen and they had to indicate only the letter b and p but not the others. The string of letters is typically so long that children cannot finish a given letter sequence before the next one appears on a new screen. The task therefore induced an element of frustration that children need to overcome. Both, the math and the stamina measures are based on an objective computer-based test (Supplementary Information, Section 1.4).

Finally, and perhaps most importantly from a policy viewpoint, we are interested in how MCII teaching affects the children's long-run school career path. For this purpose, we administered a short survey to parents in which we asked them about their child's school track in secondary school—a decision that parents must take roughly half a year before the end of primary school (grade 4). Therefore, this survey took place during the final months of primary school, i.e., about three years after the self-regulation teaching unit.

There are essentially three different secondary school tracks available in Rhineland-Palatinate, the federal state in Germany where we conducted our study: (i) an advanced track (*Gymnasium*), (ii) a mixed track (*Integrierte Gesamtschule*), and (iii) a lower track (*Realschule Plus*). In Rhineland-Palatinate, 86 percent of the children in the advanced track earn a degree that qualifies them for general university enrollment (*Abitur*), whereas only 25% percent of children in the mixed track earn this degree[55]. For children who enter the lower track in secondary school, the probability of

switching track is very small (< 5% per year)[56]. Moreover, by predetermining educational career paths, early school track choice has substantial influence on later wages[57]. Thus, the choice of the secondary school track constitutes a major educational decision that strongly affects a child's future outcomes and lifetime earnings.

Supplementary Information (SI) is provided in an online appendix that contains detailed information on participants, treatment, data collection, outcome measures, the statistical methods, the data analysis, as well as Supplementary Figures S1–S15 and Supplementary Tables S1–S15.

Data Availability Statement is provided in the Supplementary Information, Section 1.6.

Acknowledgments:

We would like to thank all teachers, schools, and educational authorities as well as all parents and children for their participation in the project. We are also thankful to countless excellent research assistants who made this field study possible. Moreover, we would like to thank Michael Wolf for support and provision of code in conducting the multiple testing correction. We are grateful for generous financial support that allowed us to conduct this project: E.F. and D.S. acknowledge support by the Jacobs Foundation (project 2013-1078-00). E.F. acknowledges support from the University Research Priority Program of the University of Zurich on Equality of Opportunity (project U-302-01-01). D.S. acknowledges support by the university research priority program "Interdisciplinary Public Policy" at Johannes Gutenberg University Mainz (project FI 2/2014-2016). H.H. acknowledges support by the German Academic Scholarship Foundation and the Research Council of Norway (FAIR, project 262675). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author Contributions:

E.F. and D.S. initiated and supervised the study throughout all stages. E.F., D.S., E.B., and K.W. conceptualized the study and all authors developed the field design. E.B., H.H., D.S., and K.W. developed intervention materials and outcome measures for the study. H.H. conducted the field experiment with input from E.B., E.F., D.S. and K.W.; E.B. and H.H. performed the data analysis with input from E.F., D.S., and K.W.; all authors were involved in the interpretation of the results and all authors wrote the paper.

Competing Interests: The authors declare no competing interests.

References and Notes:

1 Bargh, J. A., Gollwitzer, P. M. & Oettingen, G. in *Handbook of social psychology* Vol. 5 (eds Susan T. Fiske, Daniel T. Gilbert, & Gardner Lindzey) 268-316 (Wiley, 2010).

2 de Ridder, D. T. D., Lensvelt-Mulders, G., Finkenauer, C., Stok, F. M. & Baumeister, R. F. Taking Stock of Self-Control: A Meta-Analysis of How Trait Self-Control Relates to a Wide Range of Behaviors. *Personality and Social Psychology Review* **16**, 76-99, doi:10.1177/1088868311418749 (2012).

3 Duckworth, A. & Gross, J. J. Self-Control and Grit: Related but Separable Determinants of Success. *Curr Dir Psychol Sci* **23**, 319-325, doi:10.1177/0963721414541462 (2014).

4 McClelland, M. M. & Cameron, C. E. Self-Regulation in Early Childhood: Improving Conceptual Clarity and Developing Ecologically Valid Measures. *Child Development Perspectives* **6**, 136-142, doi:10.1111/j.1750-8606.2011.00191.x (2012).

5 Zhou, Q., Chen, S. H. & Main, A. Commonalities and Differences in the Research on Children's Effortful Control and Executive Function: A Call for an Integrated Model of Self-Regulation. *Child Development Perspectives* **6**, 112-121, doi:10.1111/j.1750-8606.2011.00176.x (2012).

6 Blair, C. & Raver, C. C. School readiness and self-regulation: a developmental psychobiological approach. *Annu Rev Psychol* **66**, 711-731, doi:10.1146/annurev-psych-010814-015221 (2015).

7 McClelland, M. M. & Cameron, C. E. Self-regulation and academic achievement in elementary school children. *New Dir Child Adolesc Dev* **2011**, 29-44, doi:10.1002/cd.302 (2011).

8 Moffitt, T. E. *et al.* A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 2693-2698, doi:10.1073/pnas.1010076108 (2011).

9 Heckman, J. J., Stixrud, J. & Urzua, S. The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics* **24**, 411-482, doi:10.1086/504455 (2006).

10 Bowles, S., Gintis, H. & Osborne, M. The determinants of earnings: A behavioral approach. *Journal of Economic Literature* **39**, 1137-1176, doi:10.1257/jel.39.4.1137 (2001).

11 Richmond-Rakert, L. S. *et al.* Childhood self-control forecasts the pace of midlife aging and preparedness for old age. *PNAS* **118**, 1, doi:10.1073/pnas.2010211118 (2021).

12 Duckworth, A. L., Peterson, C., Matthews, M. D. & Kelly, D. R. Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology* **92**, 1087-1101, doi:10.1037/0022-3514.92.6.1087 (2007).

13 Duckworth, A. L. *Grit - The Power of Passion and Perseverance*. (Simon & Schuster, 2016).

14 Banerjee, A. V. & Duflo, E. (Dis) Organization and Success in an Economics MOOC. *Am Econ Rev* **104**, 514-518, doi:10.1257/aer.104.5.514 (2014).

15 Ng Lee Yen, A. M. The influence of self-regulation processes on metacognition in a virtual learning environment. *Educational Studies* **46**, 1-17 (2020).

16 Huang, R. H. *et al. Guidance on Active Learning at Home during Educational Disruption - Promoting student's self-regulation skills during COVID-19 outbreak*. (Smart Learning Institute of Beijing Normal University, 2020).

17 Baumeister, R. F., Vohs, K. D. & Tice, D. M. The Strength Model of Self-Control. *Current Directions in Psychological Science* **16**, 351-355, doi:10.1177/1745691617716946 (2007).

18 Jacob, R. & Parkinson, J. The Potential for School-Based Interventions That Target Executive Function to Improve Academic Achievement: A Review. *Review of Educational Research* **85**, 512-552, doi:10.3102/0034654314561338 (2015).

19 Berkman, E. T. in *Handbook of self-regulation : research, theory, and applications* Vol. 3 (eds Kathleen D. Vohs & Roy F. Baumeister) (Guilford, 2016).
20 Baumeister, R. F. & Vohs, K. D. Strength Model of Self-Regulation as Limited Resource: Assessment, Controversies, Update. *Adv Exp Soc Psychol* **54**, 67-127 (2016).
21 Alan, S., Boneva, T. & Ertac, S. Ever Failed, Try Again, Succeed Better: Results from a Randomized Educational Intervention on Grit. *Quarterly Journal of Economics* **134**, 1121-1162, doi:10.1093/qje/qjz006 (2019).
22 Gunzenhauser, C. & Nuckles, M. Training Executive Functions to Improve Academic Achievement: Tackling Avenues to Far Transfer. *Front Psychol* **12**, 624008, doi:10.3389/fpsyg.2021.624008 (2021).
23 Santos, I. V. *et al.* Can Grit Be Taught ? Lessons from a Nationwide Field Experiment with Middle-School Students. (The World Bank, Policy Research Working Papers, 2021).
24 Pandey, A. *et al.* Effectiveness of Universal Self-regulation-Based Interventions in Children and Adolescents: A Systematic Review and Meta-analysis. *JAMA Pediatr* **172**, 566-575, doi:10.1001/jamapediatrics.2018.0232 (2018).
25 Dignath, C., Buettner, G. & Langfeldt, H. P. How can primary school students learn self-regulated learning strategies most effectively? A meta-analysis on self-regulation training programmes. *Educational Research Review* **3**, 101-129, doi:10.1016/j.edurev.2008.02.003 (2008).
26 Takacs, Z. K. & Kassai, R. The efficacy of different interventions to foster children's executive function skills: A series of meta-analyses. *Psychol Bull* **145**, 653-697, doi:10.1037/bul0000195 (2019).
27 Currie, J. Early childhood education programs. *Journal of Economic Perspectives* **15**, 213-238 (2001).
28 Diamond, A., Barnett, W. S., Thomas, J. & Munro, S. The early years - Preschool program improves cognitive control. *Science* **318**, 1387-1388 (2007).
29 Heckman, J. J. The economics, technology, and neuroscience of human capability formation. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 13250-13255, doi:10.1073/pnas.0701362104 (2007).
30 Diamond, A. & Lee, K. Interventions Shown to Aid Executive Function Development in Children 4 to 12 Years Old. *Science* **333**, 959-964, doi:10.1126/science.1204529 (2011).
31 Aizer, A. & Currie, J. The intergenerational transmission of inequality: Maternal disadvantage and health at birth. *Science* **344**, 856-861 (2014).
32 Oettingen, G. & Gollwitzer, P. M. in *Social psychological foundations of clinical psychology* (eds J. E. Maddux & J. P. Tangney) 114 - 134 (Guilford Press, 2010).
33 Oettingen, G. Future thought and behaviour change. *European Review of Social Psychology* **23**, 1-63 (2012).
34 Gollwitzer, P. M. Weakness of the will: Is a quick fix possible? *Motivation and Emotion* **38**, 305-322 (2014).
35 6, S. (MCII shares the property of general applicability to a diverse set of goals with the growth mindset approach (Dweck, C. S. and Yeaker D. S.: Mindsets - A View From Two Eras, Perspect. on Psych. Science 14(3), 481, 2019) and the GRIT approach (Duckworth A.: GRIT - The Power of Passion and Perseverance, Simon & Schuster, New York 2016). These approaches emphasize the importance of the belief that abilities are malleable and goals can be achieved with effort and perseverance. However, instilling these beliefs in young children such a first graders is extremely challenging for the same reasons that implementing MCII in this age group is challenging. In fact, MCII and GRIT are quite complementary - MCII may be considered as a metacognitive strategy that helps individuals to pursue goals with passion

and perseverance, ie., to become "grittier". Likewise, a consequence of the successful application of MCII may be that children develop a growth mindset.).
36 Wang, G., Wang, Y. & Gai, X. A Meta-Analysis of the Effects of Mental Contrasting With Implementation Intentions on Goal Attainment. *Front Psychol* **12**, 565202, doi:10.3389/fpsyg.2021.565202 (2021).
37 Gawrilow, C., Morgenroth, K., Schultz, R., Oettingen, G. & Gollwitzer, P. M. Mental contrasting with implementation intentions enhances self-regulation of goal pursuit in schoolchildren at risk for ADHD. *Motivation and Emotion* **37**, 134-145, doi:10.1007/s11031-012-9288-3 (2013).
38 Duckworth, A. L., Kirby, T. A., Gollwitzer, A. & Oettingen, G. From Fantasy to Action: Mental Contrasting With Implementation Intentions (MCII) Improves Academic Performance in Children. *Social Psychological and Personality Science* **4**, 745-753 (2013).
39 Duckworth, A. L., Milkman, K. L. & Laibson, D. Beyond Willpower: Strategies for Reducing Failures of Self-Control. *Psychological Science in the Public Interest* **19**, 102 - 129 (2019).
40 Frison, L. & Pocock, S. J. Repeated Measures in Clinical-Trials - Analysis Using Mean Summary Statistics and Its Implications for Design. *Statistics in Medicine* **11**, 1685-1704, doi:DOI 10.1002/sim.4780111304 (1992).
41 McKenzie, D. Beyond baseline and follow-up: The case for more T in experiments. *Journal of Development Economics* **99**, 210-221, doi:10.1016/j.jdeveco.2012.01.002 (2012).
42 52, S. (In addition, we also estimate our treatment effects using Tobit models to account for censored outcome variables; all results are robust to this alternative specification (see SI, Tables S9 and S10). The results also do not change if we restrict the sample to those observations that are present in all four waves (no attrition sample, SI, Tables S11-S13)).
43 Dynarski, S. M. *For Better Learning in College Lectures, Lay Down the Laptop and Pick up the Pen*. (The Brookings Institution, 2017).
44 Slavin, R. E. Best-evidence synthesis: an alternative to meta-analytic and traditional reviews. *Education Research* **15**, 5-11 (1986).
45 Hill, C. J., Bloom, H. S., Black, A. R. & Lipsey, M. W. Empirical Benchmarks for Interpreting Effect Sizes in Research. *Child Development Perspectives* **2**, 172-177, doi:DOI 10.1111/j.1750-8606.2008.00061.x (2008).
46 Kraft, M. A. Interpreting Effect Sizes of Education Interventions. *Educational Researcher* **49**, 241-253 (2020).
47 Hanushek, E. A. & Rivkin, S. G. in *Handbook of the Economics of Education* Vol. 2 (eds E. A. Hanushek & F. Welch) Ch. 18, 1051 - 1078 (North Holland, 2006).
48 Yeager, D. S. *et al.* A national experiment reveals where a growth mindset improves achievement. *Nature* **573**, 364-+, doi:10.1038/s41586-019-1466-y (2019).
49 Bailey, D., Duncan, G. J., Odgers, C. L. & Yu, W. Persistence and Fadeout in the Impacts of Child and Adolescent Interventions. *J Res Educ Eff* **10**, 7-39, doi:10.1080/19345747.2016.1232459 (2017).
50 Cunha, F. & Heckman, J. The technology of skill formation. *Am Econ Rev* **97**, 31-47, doi:10.1257/aer.97.2.31 (2007).
51 Berger, E. M., Fehr, E., Hermes, H., Schunk, D. & Winkel, K. *The Impact of Working Memory Training on Children's Cognitive and Noncognitive Skills* (University of Mainz, 2022).
52 7, S. (In German a ball is masculine. Therefore, Hurdy's was a "he" in our story).
53 51, S. (The last sentence in this story is abbreviated and the full story reads as follows: "He rolled to his uncle's garage. He looked around for a while and then found a dripping bottle of glue, some wire, and old gloves in some shelves in a corner. He waited under the bottle of

glue until a drop of glue fell on him. He rolled cleverly the exact distance until the wire stuck to him. He let the glue dry a bit, and the little ball already had a first arm. He stuck this arm into a glove. He repeated the process on the other side, and he then had two arms and two hands. He then found two springs in an old, broken mattress and a pair of sneakers. He used the springs as legs, and then put the shoes on. With a happy feeling, the little ball went on his way to the mountain.").

54 Cohors-Fresenborg, E., Kramer, S., Pundsack, F., Sjuts, J. & Sommer, N. The role of metacognitive monitoring in explaining differences in mathematics achievement. *ZDM* **42**, 231-244 (2010).

55 Rhineland-Palatine, S. O. *Allgemeinbildende Schulen im Schuljahr 2017/2018*. (Statistisches Landesamt Rheinland-Pfalz, 2018).

56 Bellenberg, G. Schulformwechsel in Deutschland. Durchlässigkeit und Selektion in den 16. Schulsystemen der Bundesländer innerhalb der Sekundarstufe I. (Bertelsmann Stiftung, 2012).

57 Dustmann, C. Parental background, secondary school track choice, and wages. *Oxford Economic Papers-New Series* **56**, 209-230, doi:10.1093/oep/gpf048 (2004).

Supplementary Information

(SI)

for

Teaching Self-regulation

Daniel Schunk, Eva M. Berger, Henning Hermes, Kirsten Winkel, Ernst Fehr

Contents

1 Supplementary Methods

The study was conducted in primary schools in Mainz, Germany in 2013/2014. It consisted of a five-week intervention, four data collection waves, and a long-term follow-up survey three years after the intervention. Our study received ethical approval from the Human Subjects Committee of the Faculty of Economics, Business Administration and Information Technology at the University of Zurich in September 2012. We confirm that we have complied with all relevant ethical regulations.

The study consisted of a pre-intervention data collection wave (t_0), the five-week intervention period, a data collection wave shortly (4–5 weeks) after the intervention (t_1), and two follow-up data collection waves 6 and 12–13 months after the intervention (t_2 and t_3).

We provide details on participants (Section 1.1), the treatment condition (Section 1.2), the data collection waves (Section 1.3), the outcome measures used (Section 1.4), the data analysis (Section 1.5), and other details (Section 1.6) below. Supplementary figures are provided in Section 2, and all supplementary tables in Section 3.

1.1 Supplementary Details on Participants

Sampling of Participants

In February 2012, we received the approval from the Federal Ministry for Education in Rhineland-Palatine to conduct the study with first graders in the city of Mainz. The authority responsible for elementary schools in Mainz (ADD) contacted schools and provided us with a list of elementary schools in May 2012. We selected 12 schools for participation in the study based on two criteria: being located in the city of Mainz and the possibility of including at least two school classes per school in the study. The participating schools agreed that (i) over a period of five weeks regular schooling lessons would be replaced by a self-regulation teaching unit and that (ii) the children would participate in all four planned data collection waves.

Final Sample and Attrition

We recruited 12 schools with 31 classes for the study. The sample consisted of three schools with four classes, one school with three classes, and eight schools with two classes. There were 599 children in these classes in November 2012. We received 580 parental consent forms that allowed us to collect data in evaluation waves t_0–t_3, resulting in a consent rate of 96.8%. We were able to evaluate 572 children of the 580 for whom we received parental consent to collect data for our final data set. The children we could not evaluate either switched to non-participating classes or schools, moved away, or were ill for a longer period of time during data collection; we did not exclude any available data. Among the sample of 572 children, 292 were girls (51%) and 280 were boys (49%). Mean age prior to the intervention (Jan 2013) was 6.84 years (SD = 0.36 years).

Our sample decreased from 572 children in t_0 (at baseline) to 531 children in t_3 (12–13 months after treatment) due to attrition (see Figure S1). This corresponds to an attrition rate of 7.2%. This attrition was due to children who switched to non-participating classes or schools, moved away, or were ill for a longer period of time during data collection; we did not exclude any available data. Attrition appears non-selective, as we find that the estimated treatment effects remain stable when we restrict the sample to only those children who remain in the sample throughout all waves. Results for these estimations can be found in Tables S12–S14.

We also tried to conduct another randomized field study in Switzerland but failed to do so because the relevant school authorities were not able to ensure randomization of school classes into treatment and control classes: several schools/classes were only willing to participate under the condition of being assigned to the control group.

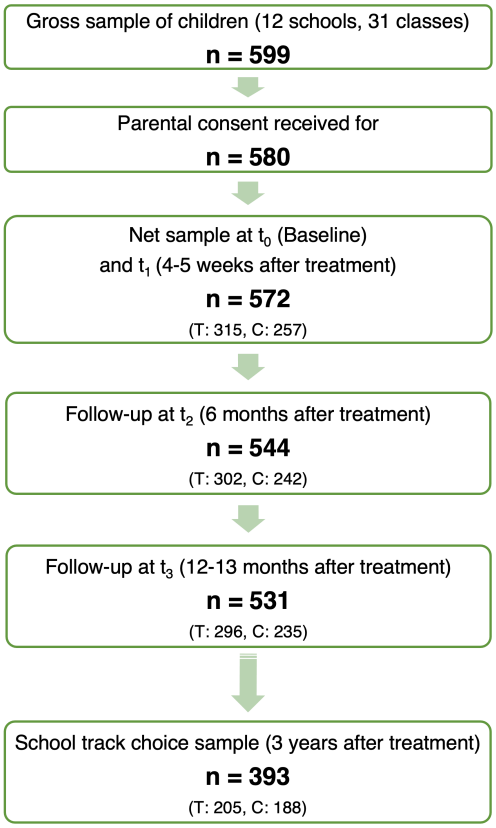


Figure S1: Flow of Participants through the Study (T: Treatment, C: Control)

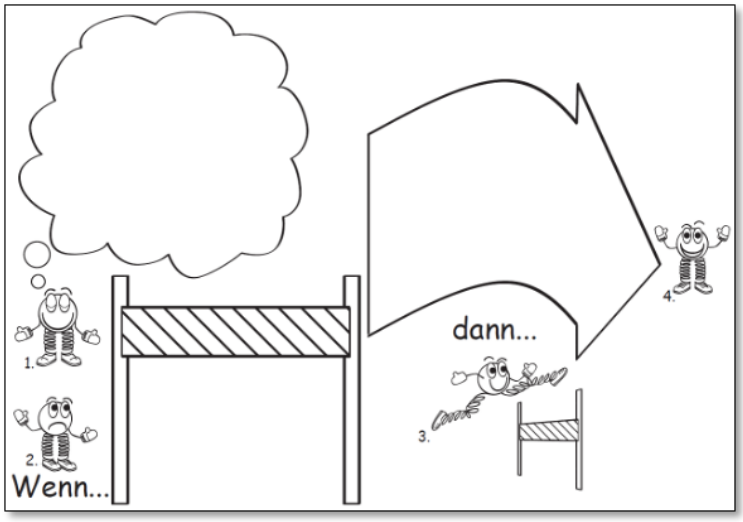


Figure S2: Example of a Workbook Page to be Filled Out By a Child During the Teaching Unit

1.2 Supplementary Details on the Self-regulation Teaching Unit

Background of Self-regulation Strategy and Implementation for First Graders

We developed a teaching unit for the treatment condition that was based on a metacognitive self-regulation strategy called "Mental Contrasting with Implementation Intentions" (MCII). MCII consists of two separable concepts from motivational psychology, namely "mental contrasting" and "implementation intentions". The first process, i.e., mentally contrasting a positive future with the negative present reality, aims at producing goal commitment (depending on levels of expectations about success) and supporting goal striving [1]. The second process, implementation intentions, addresses the challenge of overcoming the major barriers to goal attainment, e.g., getting started, not getting distracted, doing goal-directed activities first, overcoming barriers, etc. [2]. Implementation intentions usually take the form of a "when-then-plan", e.g., "**When** situation Y occurs, **then** I will respond with (goal-directed) action Z!" and have been shown to improve the attainment of a wide range of goals [3]. Empirical evidence suggests that combining the two concepts and teaching people to apply this MCII strategy is effective in improving outcomes across a range of domains such as health and nutrition behaviors [4].

However, as pointed out in the main text, no study has so far integrated MCII-based teaching lessons into normal school teaching. Two previous studies recruited sixth- and seventh graders and fifth-graders, but MCII was provided by trained experts outside of regular classroom teaching in both studies [5, 6]. Thus, the concept has not yet been adapted for first graders with very limited reading and writing skills. In addition, children at that age are often characterized by limited goal setting skills, patience, attention span and inhibition skills, as well as a lack of perseverance which may pose major challenges for integrating MCII-based self-regulation into the normal school curriculum. In order to address these challenges, we translated the strategy into several child-oriented and easy-to-recall steps. We made the steps vivid and meaningful for the children by developing a role model ("Hurdy") and a story that served as a metaphor for the single steps of the strategy (See Figures 1 and 2 in the main text).

We developed five teaching lessons on the basis of extensive and detailed instructional material for the teachers and their students such that MCII-based self-regulation could be taught in a similar, standardized way across classes and teachers. The material consisted of a teacher's manual, an illustrated storybook, a classroom poster, some flashcards, a stamp and a workbook for the students. The material is based on a main character (named "Hurdy",

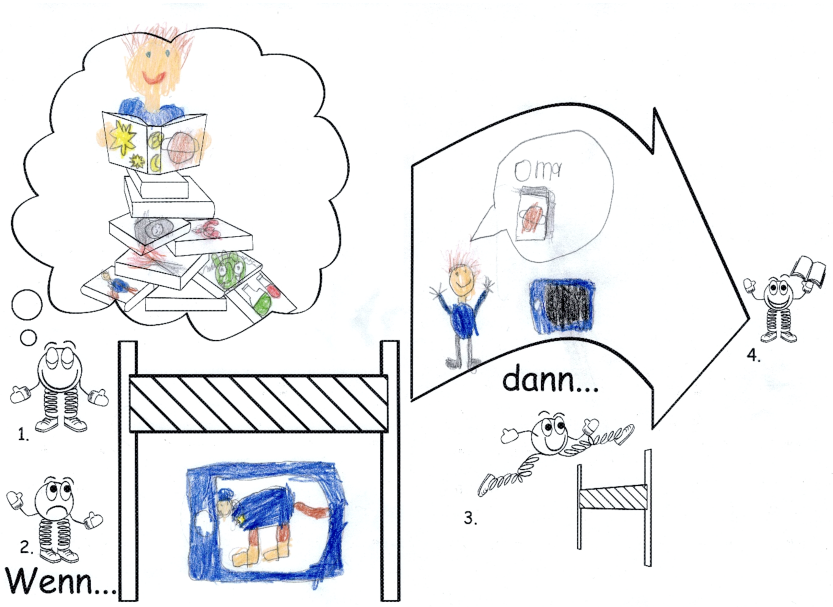


Figure S3: MCII-strategy applied to the first goal in a boy's workbook.

Notes: 1. The child imagines the future outcomes of becoming better in reading and draws books he wants to read about the universe, rockets, volcanos, experiments and favorite stories. 2. The child contrasts the desired outcome to his typical obstacle (e.g., the temptation of using his tablet). 3. The child develops a goal-oriented action to overcome his obstacle. 4. The child forms an implementation intention (e.g., "When I see my tablet, then I first call my Grandma to read aloud to her.").

derived from "hurdle jumper"). As the students just began to learn reading and writing before the intervention but had not yet learned all letters of the alphabet, one of the key challenges was to teach the MCII strategy without requiring any writing or reading skills. Thus, we put the main idea of the strategy into an illustrated storybook which the teacher read out for the students step by step. In this story, each step of the abstract MCII strategy was translated into a concrete activity of the energetic and emotionally appealing main character of the story (see Fig. 1 in the main text). With this main character's activity in mind, it was easy for the first graders to identify with the main character's wishes, his obstacles, and his when-then-plans used to pursue his goals. Besides the illustrated storybook, the teachers received a detailed manual to implement each lesson precisely according to the suggested timetable and instructions for each lesson. The children used a prepared workbook to develop personalized strategies (see Fig. S2 for an example page in students' workbook) and a stamp was used for giving individual feedback to achieved goals in the workbook. In addition, a large poster similar to the workbook page in Fig. S2 (but without the children's added drawings) was on the wall in each classroom and related flash cards attached to the poster (see Fig. S4) were used for demonstrations.

The self-regulation teaching unit consisted of five focused teaching lessons of 50 min each that were evenly spread over the course of five weeks. Teaching lessons were conducted by the regular classroom teacher. Prior to the intervention, these teachers participated in a three-hour training workshop, during which we taught them (i) the theoretical idea of the MCII strategy, (ii) the concept for the practical implementation, and provided them with (iii) detailed instructions for the teaching material, which we developed and designed for the study (see above). In each teaching lesson a trained research assistant observed the lessons and recorded a comprehensive documentation of the training as well as compliance to the teaching protocol in an unobtrusive way; in all classes, compliance was very high. During the course of the self-regulation teaching unit, children pursued three different goals: First, get better in reading; second, detect own mistakes by using a monitoring strategy (see Fig. S5); third, individual goals of personal importance.

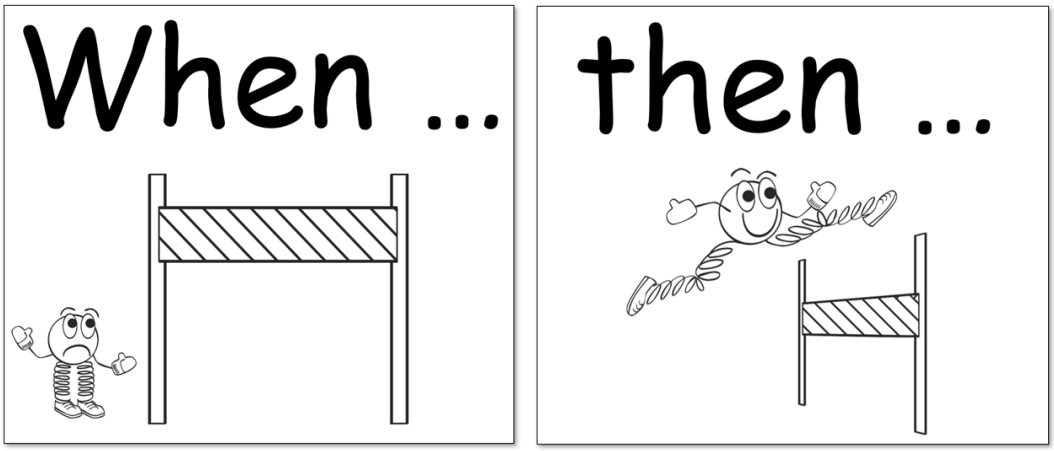


Figure S4: Example for Two Flashcards Used to Develop When-then-Plans (here: translated from German)

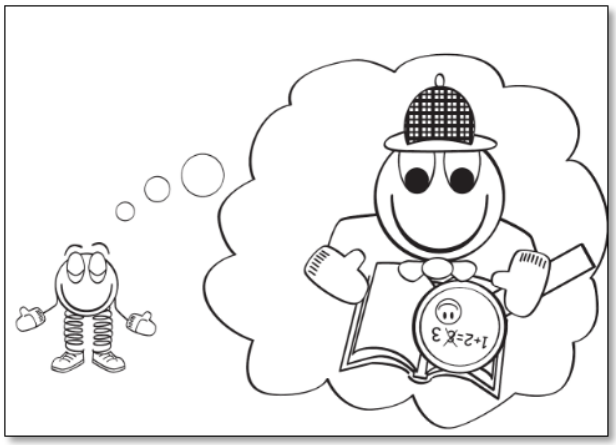


Figure S5: Example for Introducing the Second Goal (Careless Mistakes) in the Storybook

1.3 Supplementary Details on the Data Collection

The main data was collected at four points in time: the first wave took place immediately before the intervention (t_0), the second wave took place 4–5 weeks after the intervention (t_1), the third wave took place six months after the intervention (t_2), and the fourth wave took place 12–13 months after the intervention (t_3). In each wave, we collected computer-based outcomes that measured self-regulation abilities and academic achievement. We describe these outcome measures in detail below. In addition, we administered questionnaires to teachers and parents. In t_3, we also asked the children a few questions after the computer-based tests.

The data collection was run by a professional data collection service provider experienced with conducting research projects in these settings. The tests were conducted outside the classroom; both the children from the control and from the treatment groups participated in the tests. The data collection was conducted by interviewers experienced in standardized testing procedures and in working with children of that age. They were trained in an eight-hour training session run by the data collection service provider together with the authors of this study. Importantly, the interviewers involved in administering the tests to the children (i.e., the employees of the data collection service provider) were blind to the children's assignment to the treatment conditions. The teachers were not involved in the design and the conduct of the tests, and they did not even know the content of the tests, i.e., it was impossible for the teachers to prepare the children for the tests. Finally, three years after the treatment, we also conducted a survey on school track choice, details see below. This study reports all measures in this project up to and including the survey on school track choice.

Testing Procedures

The tests were administered using computers with 22" touchscreens and headphones. The instructions were auditive via headphones and supported by visual demonstrations shown on the screens. The children entered their responses using touchscreens that were easy to handle.

The tests were run in two blocks of about 30 minutes, scheduled on two consecutive days, primarily during the first or second lesson of the school day. Tests were done in groups of five children supervised by one "interviewer". Each child sat in front of a touchscreen positioned in a standardized way on the desk and had headphones to listen to the instructions. All children started at the same time, but could complete the test at their own pace. The whole testing procedure for a class lasted for about three to four school days. Due to these testing procedures we achieved an exceptionally high degree of standardization, especially through the instructions via headphones.

All tests were pretested in another primary school that did not participate in the study. All children received a small toy for participating in the evaluation wave. Over the four data collection waves, the tasks became generally more difficult to account for the increase in children's abilities over time.

Parent Questionnaires

Parent questionnaires were only distributed in the data collection waves t_0 and t_2, i.e., before the intervention and 6 months after the intervention. Parent questionnaires included questions on socio-demographic characteristics of the family, parental behavior (also towards the child as well as educational goals) and parental characteristics as well as the child's personality, attitude towards school, general health, and everyday behavior (including SDQ). Parents filled out 467 out of 572 parental questionnaires in t_0 (82%) and 419 out of 544 in t_2 (77%).

Teacher Questionnaires

In each data collection wave, teachers filled out a questionnaire. These questionnaires contained questions on children's characteristics and behaviors, and teacher characteristics and behaviors, as well as experience with and expectations about the intervention (if they were in the treatment group). In particular, we asked the teachers in every data collection wave to assess each child's self-regulation abilities using several questions (see Section 1.4). We achieved a 100% return rate for the teacher questionnaire in all four evaluation waves.

Survey on Secondary School Track Choice

In addition to the main data collection, we administered a short survey to parents and teachers when children were in the final grade of primary school (grade 4). This survey was conducted in April 2016, i.e., three years after the treatment, and asked parents about the secondary school track the child was enrolled for grade 5. The questionnaire was sent to participating schools and teachers distributed and collected questionnaires. Parents submitted their answers in a sealed envelope, so that the teacher could not see their response. Teachers also provided a recommendation which school track the child should attend. However, in our study context the school track decision is taken by the parents, and teachers' recommendation is not binding for the children.

We received a total of 393 questionnaires (74% of the sample in t_3). This attrition was due to reasons such as children moving away from the city of our study or parents not answering our follow-up questionnaire. Importantly, attrition is not systematic: If we regress participation in the survey on secondary school track choice on the treatment condition, gender of the child, age of the child, and school fixed effects we do not find any significant treatment effect regardless of whether we use a linear probability regression or a probit regression. Thus, we conclude that there was no significant difference in attrition between treatment and control

group. However, to be on the safe side, we nevertheless control for any residual nonsignificant differences in attrition by applying inverse probability weighing when we analyze the impact of self-regulation teaching on secondary school track choice.

1.4 Supplementary Details on Outcome Measures

This section describes the tests that we used to measure the effects of the self-regulation teaching unit. For the assessment of treatment-related outcomes, we use a computer-based reading test and teacher-ratings of overall reading ability as well as whether children commit many careless mistakes. For evaluating improvements on self-regulation abilities in a broader range, we use computer-based tests of inhibition and attention as well as teacher-ratings on six items related to self-regulatory behavior (similar to "behavioral grades"). Finally, we report results on untrained skills, namely math abilities and a letter discrimination task ("bp task"). We also measured children's working memory capacity, fluid IQ, children's reading habits, and time and risk preferences using computer-based and non-computer-based tasks, but these measures are not part of the present study. For the ease of interpretation and comparison, we standardize all test scores to mean = 0 and SD = 1, separately by test and wave. Histograms of the distribution of all raw test scores (i.e., before standardization) for t_0 and t_3 are displayed in Fig. S14.

Self-regulation Teaching Outcomes

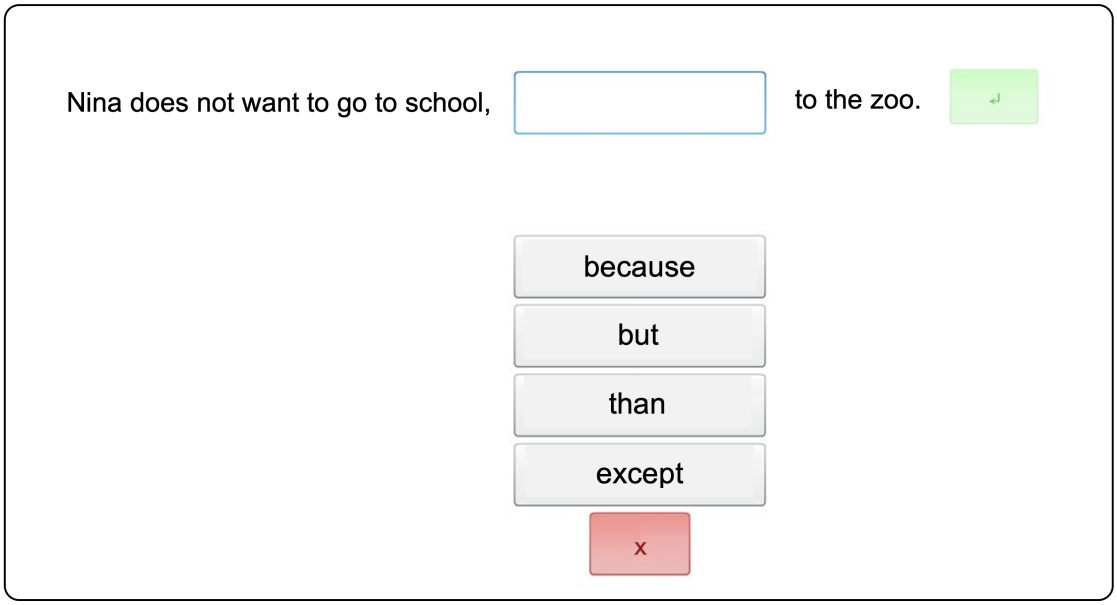
Reading Comprehension Skills: Reading comprehension was assessed by a sentence comprehension test in single choice format. On the screen (see Fig. S6), a sentence with one gap was presented in a line. To fill the gap, the children had to choose from a list of four alternatives presented below the gap. Tapping on one of the words in the list made it appear in the gap. Children could correct their choice by using the red X button below the list. Children had to confirm their choice by pushing the green enter-button right beside the sentence.

Generally, there was only one word missing in the sentence. In t_2 and t_3 there were also a few gaps to be filled with a combination of two short words. The difficulty of the items was multidimensional. It varied within a test, and in particular between the evaluation waves, where it was adjusted to the curriculum. In t_0 and t_1, the test contained 10 sentences consisting of 3 to 9 words per sentence. The words only contained those letters that had already been introduced to the children in earlier lessons during the school year. As most children become much faster in reading before t_2, the reading comprehension task contained 16 sentences with 4 to 15 words per sentence in t_2, and 16 sentences with 4 to 16 words per sentence in t_3.

Overall Reading Ability: The above-presented measure for reading abilities stems from an objective, computer-based test and not on subjective assessments. However, these types of test can, at the same time, only capture specific aspects of overall reading ability. To get a broader picture of children's ability in the area of reading, we additionally asked teachers in the teacher questionnaire (see Section 1.3) for their assessment of children's overall reading ability. In each wave, teachers evaluated each child's overall reading ability using the following statement: "Overall, the child shows good reading achievements." Teachers answered on a 7-point Likert-type scale with 1 = "does not apply at all" and 7 = "fully applies". We standardize the score to mean = 0 and SD = 1 within each wave.

We are aware of the fact that the teachers' subjective assessments may be a less reliable measure than objective tests. However, if the treatment affects the objective test measure and the subjective measure in similar ways, our confidence in the reliability of the treatment effect is strengthened.

Careless Mistakes: In order to measure whether the self-regulation teaching module on committing less careless mistakes was successful, we also use information from the teacher



Example for easy item:

Leo is at the [ ].
(answer options: mum, lake, hat, name)

Example for difficult item:

In good weather, Fabian takes the bike [ ] he better likes to go by foot in bad weather.
(answer options: while, during, as if, without)

Figure S6: Reading Comprehension Task, Screenshot Plus Two Further Examples

questionnaires (see Section 1.3). In each wave, teachers rated each child on the statement "The child makes a large number of careless mistakes.", using a 7-point Likert-type scale with 1 = "does not apply at all" and 7 = "fully applies". We reverse-code the scale (such that higher values refer to better outcomes) and standardize the score to mean = 0 and SD = 1 within each wave.

Self-regulation Abilities

Go/No-Go Task: To measure inhibitory abilities, we employed a go/no-go task that was adapted from Gawrilow & Gollwitzer [7]. In this task, the child had to push a red button on the touchscreen every time one of four different animals appeared on the screen (rooster, mouse, cat, pig—see Fig. S7 below). However, the children were told not to push the red button for one other animal (cow). The procedure of the task is as follows: The red button is displayed on the touch screen throughout the task. In addition, the children first see an X in the middle of the screen for 0.6–1.2 seconds (these times randomly vary across items but are equal across waves). Then the picture of an animal appears with a display time of 1.55 seconds and a time slot for reaction of 1.55 seconds (the display time for the animal was reduced to 0.65 seconds in t_1, t_2, and t_3.) In this time window, the children must decide whether to push the button and to implement the button press. Subsequently, the children again see the X, then the picture, and so on. In total, 50, 60, 70, and 80 items were presented in t_0, t_1, t_2, and t_3, respectively. In t_0 and t_2, the pictures were animals as described above. The pictures were vehicles in t_1 and t_3 (go = car, train, ship, airplane; no-go = truck).

We measure performance in this task in two ways. First, we compute the commission errors (i.e., the number of times a child fails to inhibit the "go-response" when a no-go item is displayed), multiply by -1, and standardize the score to mean = 0 and SD = 1 within each

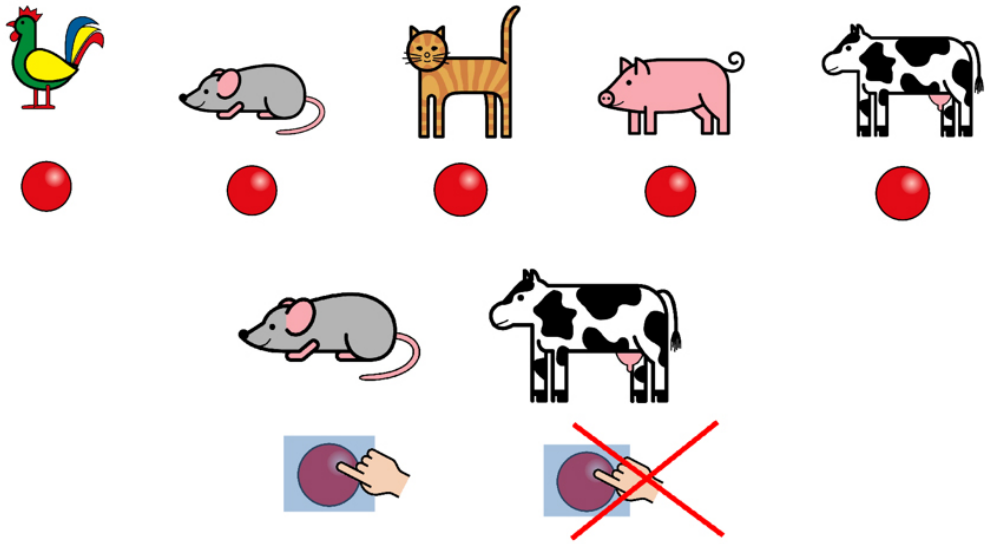


Figure S7: The Animals and the "Go-Button" in the Go/No-Go Task

wave. Thus, a higher score indicates better performance in the task (i.e., fewer mistakes). This outcome is used to assess inhibition control, i.e., the ability to inhibit unwanted, automatic responses and impulsive actions.

Second, we analyze the omission errors (i.e., the number of times a child fails to push the red button when a go item is displayed), multiply by -1, and standardize the score to mean = 0 and SD = 1 within each wave. Hence, a higher score indicates better performance in the task (i.e., fewer mistakes). We use this outcome to measure attention, i.e., the ability to focus on a given task and not to be distracted.

Overall Self-regulation: Similar to how we measure reading abilities, we wanted to complement the objective, computer-based go/no-go task with a broader assessment of self-regulation abilities in everyday classroom behavior, using the teacher questionnaire. As mentioned above, we are aware of the fact that the teachers' subjective assessments may be a less reliable measure than objective tests. However, if the self-regulation teaching affects the objective test measure and the subjective measure in similar ways, our confidence in the reliability of the treatment effect is strengthened.

In each wave, teachers assessed the self-regulation abilities of each child in their class by answering the questions listed below in each data collection wave. Questions 1–4 were answered by means of a 7-point Likert-type scale with 1 = "does not apply at all" and 7 = "fully applies". The answer options for the questions 5 and 6 are indicated below.

1. The child works in a concentrated and enduring manner.
2. The child makes a large number of mistakes due to inattention (reverse coded).
3. The child has a lot of self-discipline.
4. The child has trouble waiting for his/her turn (reverse coded).
5. The child disturbs class instruction often (reverse coded).
6. Please indicate for each child how often he/she forgot his/her homework or did not do his/her homework despite having an assignment in the last six months? (1 = never forgot homework up to 7 = forgot homework often) (reverse coded)
7. How do you rate the child with respect to patience? (1 = very impatient, 7 = very patient)

In this study, we analyze item number 2 ("The child makes a large number of mistakes due to inattention", reverse coded) separately because this skill was the second goal of our intervention (see Section "Addressing the challenges of teaching MCII to first graders" in the main paper). The remaining six items above are analyzed as overall self-regulation. Note that if we combine item number 2 with the remaining six items to obtain an alternative construct for overall self-regulation, then our treatment effects in t_1, t_2, and t_3 correspond to 0.24 ($p = 0.002$), 0.35 ($p = 0.005$), and 0.619 ($p < 0.001$), i.e., the conclusions from the paper do not change.

Importantly, the teachers had to answer these questions subsequently for each child in their classroom, i.e., they first rated all children in their class on item 1, then all children in their class on item 2, etc. This makes it very unlikely that item correlations within child (or improvements on several dimensions for a child) were driven by teachers just using the same point on the scale for all items for a single child (e.g., by seeing child A's name and then clicking on a "6" for each item for this child—this was not possible in our survey design).

The items above were developed with the purpose of assessing young children's self-regulation skills in a classroom context. Also, as many other studies look at effects of self-regulation interventions on "behavioral grades" or "behavioral conduct", we wanted to have information on these dimensions for children in our sample; however, at this age there are no such "grades" or remarks on bad conduct available from the teachers or schools. With this in mind, we designed the questionnaire, using (and adapting) items from the Strengths and Difficulties Questionnaire (SDQ) proposed by Goodman [8], and on the Self-Control Scale developed by Tangney *et al.* [9], which was translated into German and validated by Bertrams & Dickhäuser [10].

The go/no-go task measures children's ability in inhibitory control and attention using an objective performance measure that does not depend on an evaluator's subjective impressions. In contrast, teachers' assessments of children's self-regulation abilities rely on both teachers' actual experiences with the children and their subjective interpretation of these experiences. Thus, a sceptic might put less weight on the overall self-regulation score, but, in principle, one can lend support to the subjective measure by examining whether it correlates with the objective performance measure. This is based on the idea that if teachers' assessments contain an objective rationale, i.e., that their assessments have a meaningful objective basis and are not purely subjective impressions, then we should observe a positive correlation.

We therefore computed the correlation between the objectively measured performance in the go/no-go task (averaged for each child over t_0–t_3) and the teacher-rated assessment of children's self-regulation scores (again averaged over t_0–t_3) to assess the credibility of teachers' ratings. We find highly significant correlations of $\rho = 0.36$ ($p < 0.001$) for inhibition, and $\rho = 0.32$ ($p < 0.001$) for attention. These correlations are at least in line or even larger in size than what has been found in other studies [11, 12]. Thus, it appears that overall self-regulation does indeed contain useful information about children's self-regulation abilities.

Parental Ratings of Self-regulation: To further examine the impact of the self-regulation teaching unit on "overall self-regulation", we also analyze parental ratings of items which are related to self-regulatory skills. However, the parental data pose several challenges that make it more difficult to analyze and interpret this data: (i) parental ratings are only available at baseline (t_0) and for the 6-months follow-up (t_2), (ii) the sample size is substantially smaller (for details, see below), and (iii) parental ratings suffer from several problems adding noise to the measurement (e.g., parents experience children in a different context (home vs. school), they usually do not have the possibilities to compare their child with many other children (like teachers do), and children might behave very differently in the school context compared to the home context).

As noted in Section 1.3, we received n = 467 parental questionnaires in t_0 and n = 419 questionnaires in t_2. Since we control for the baseline values of our outcomes in the analysis, the effective sample is reduced to n = 386 parents responding in t_0 *and* t_2. Because parental questionnaires were administered in paper-pencil format, we also observe some item non-response. Therefore, our final sample size for this analysis is n = 363 (which corresponds to

Figure S8: The Input Device for the Arithmetic and Geometry Tasks

63% of our sample at baseline).

In the parental questionnaire, there are five items that enable us to construct a measure for parent-rated overall self-regulation:

1. My child completes tasks, can concentrate for long periods of time.
2. My child has a lot of self-discipline.
3. My child has trouble waiting for his/her turn (reverse coded).
4. My child frequently interrupts me or other adults during conversations (reverse coded).
5. How do you rate your child with respect to patience? (1 = very impatient, 7 = very patient)

The parent-rated index of overall self-regulation is computed using the exact same procedure as for the teacher-rated items, i.e., by adding up the standardized scores of all five items and then standardizing the sum again to mean = 0 and SD = 1. We then use our standard OLS model and regress the level of parent-rated overall self-regulation in t_2 on the treatment dummy, parent-rated overall self-regulation in t_0, and our standard set of controls (including school-fixed effects, gender and age). Results are reported in Table S15.

Transfer Outcomes

Math Abilities:

Arithmetic skills were assessed using three different subtasks: a number sense task, an auditory arithmetic task, and a written arithmetic task. The children had to infer/compute a correct number from the presented stimuli in all three arithmetic tasks. Children had to enter the number in an input device on the computer screen that looked like a pocket calculator (see Fig. S8 below). For example, if the child thought that the correct number is '23' she had to tap first a '2' so that this number appeared in the empty top left rectangle of the device; then she had to tap on the number '3' on the input device so that the number 23 appeared in the top left rectangle of the device. If the child was satisfied with her answer, she had to confirm by tapping on the green arrow on the top right corner. If the child wanted to correct her answer, she could do so by tapping on the red "X" on the bottom left corner of the input device.

Note that the children also had to identify a correct number in the geometry task described below, again using the same input screen in that task.

Number sense task. In this subtask, the children were presented a number of balls on a two by ten grid that was only shown for 1.7 seconds (see Fig. S9 below showing several different examples with various levels of difficulty). In general, the display time was too short to count

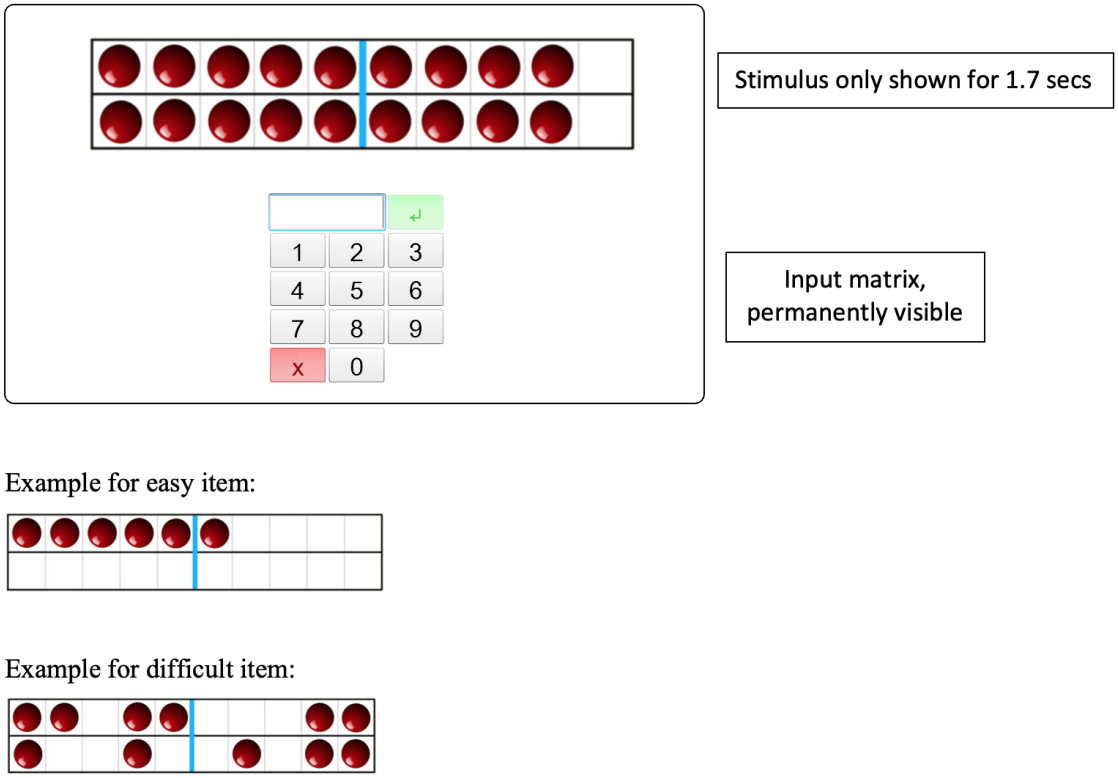


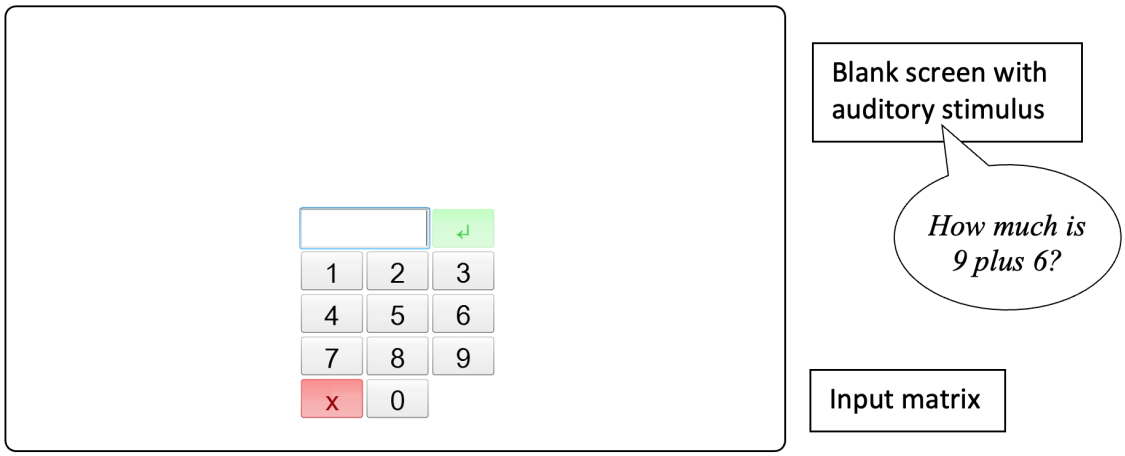
Figure S9: Number Sense Task, Screenshot Plus Two Further Examples

all balls before they disappeared. After the grid had disappeared, the children had to type the correct number of balls in the grid.

A two by ten grid with the subdivision at 5 is used in the first grade in the participating primary schools to teach numbers and calculations. To solve the number sense task, children need to be familiar with the number range up to 20, and a good understanding of the logic of the grid is useful. Because the children could not count the balls due to the short display time, they had to capture the pattern of the balls. This involves the assessment of structures as well as the detection of possible subgroups and the number of balls per subgroup. Children had to sum up the number of balls from different subgroups or use subtraction in cases where only a few balls were missing in the grid. For example, consider the first grid below (see Fig. S9) with 18 balls: Depending on the child's mathematical experience, different strategies are possible in this grid. A child knowing that 20 balls would fit in the grid and noticing that 2 balls are missing at the right end of the grid could compute 20-2=18 to arrive at the correct solution. Another child might recognize 10 balls (2 rows with 5 balls each) in the left half and 8 balls (2 rows with 4 balls each) in the right half of the grid. This child will reach the correct solution by mentally computing 10+8 after the balls have disappeared. The third grid below (see Fig. S9) gives an example of a rather difficult item. Children had to quickly recognize and structure four groups of balls containing different numbers of balls each. The children had to capture the number of balls in each subgroup simultaneously and to correctly sum up 3+3+1+4. As one of the fundamental steps in mathematical development at this age is to replace counting strategies by computing strategies, it is important that the display time was too short to be able to count the balls.

The number of balls and their distribution within the grid varied across the items and evaluation waves and was adjusted to the development of children's mathematical skills. The size of the grid, however, remained constant over time.

Auditory arithmetic task. This subtask measures arithmetic skills for addition and subtraction of two numbers (see Fig. S10). Computational tasks were presented over the headphone



Example for easy item:
"How much is 2 plus 5?"

Example for difficult item:
"How much is 92 minus 17?"

Figure S10: Auditory Arithmetic Task, Screenshot Plus Two Further Examples

(e.g. "How much is 9 plus 6?"). Children had to enter their answer into the input matrix. Each item in this task contained two numbers to be added or subtracted. Each evaluation wave contained 10 of these auditory arithmetic items.

The difficulty level was adapted to the school curriculum, e.g., with regard to the number range: In t_0 and t_1 the number range was up to 20, while in t_2 and t_3 it expanded to 100. Other major changes across waves are the increase in complexity of the mental operations and the need for numerical comprehension. Moreover, for the more difficult items, such as "92 minus 17", children needed to compute intermediate steps: First, many children would compute 92 minus 10 and keep the intermediate result 82 in mind. Then, they would subtract the remaining 7 from 82, leading to the final result.

Written arithmetic task. In contrast to the auditory task, the arithmetic problems in this subtask were not presented over the headphones but displayed on the screen. Most problems contained more than two numbers that needed to be added or subtracted; the reason for this is that we tried to avoid having children draw a result from their longer-term memory without computing. Each arithmetic problem was visible on the screen during the whole trial (see Fig. S11). Because of this (i.e., because the subjects did not need to recall the numbers from memory), the difficulty level of the required mathematical operations was generally set to be higher than in the auditory task. Children were, for example, required to add and/or subtract three or four numbers. The difficulty level was also adapted to the curriculum, analogously to the way it was done in the auditory arithmetic task.

Geometry skills. Geometry skills were assessed by a test that required the children to assess how many simple-shaped objects—such as triangles, squares, or rectangles—fit into a larger geometric object (see Fig. S12 below). Depending on the size and the shape of the larger geometric object, this task can be made harder or easier.

The task contained 10 items in each evaluation wave. The difficulty level varied across items and evaluation waves. Difficulty varied along various dimensions. Consider the easy item shown in Fig. S12 (the red square): children could solve the problem without any mental

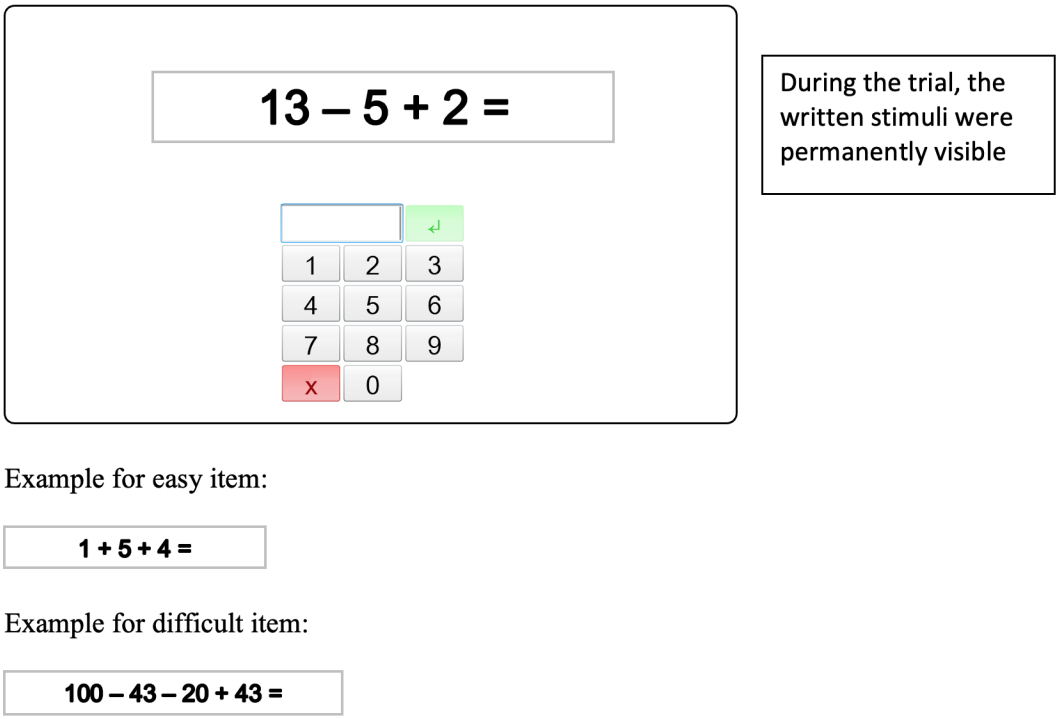
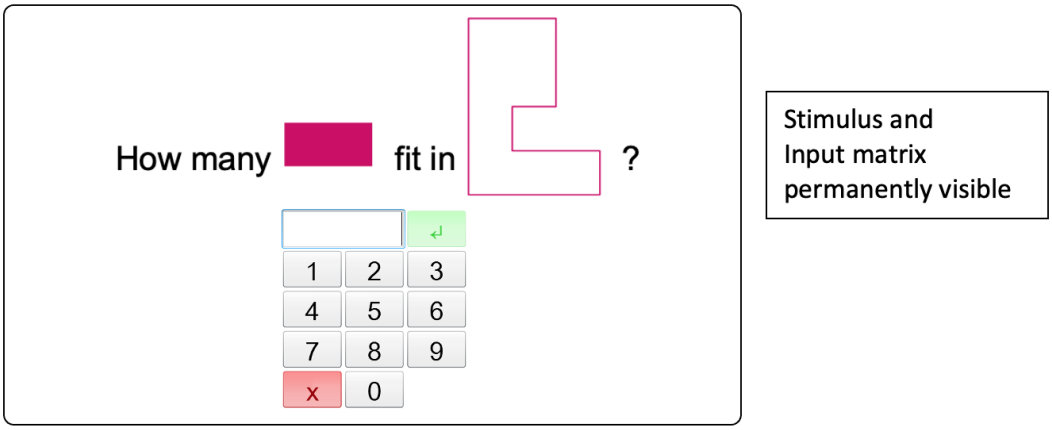


Figure S11: Written Arithmetic Task, Screenshot Plus Two Further Examples

rotation of the small square. Furthermore, the larger object is subdivided into two components, making the task even easier. In contrast, for the first item shown in Fig. S12 (the pink rectangle), children had to mentally rotate the small object to solve the question. For the difficult item in Fig. S12 (the green triangle), children hat to mentally rotate the triangle, store the number for subparts and keep track of which parts were already counted when filling the larger geometric object.

Computation of final math test score. For each of the four subtasks (number sense, auditory and written arithmetic tasks, and geometry), we added up the number of correctly solved items and standardized each subtask score to mean = 0 and SD = 1 within each wave. We then added up the four standardized subscores and standardized this composite score to mean = 0 and SD = 1 to achieve comparability to the other test scores used in our analysis. We increased the difficulty of these tasks across the four evaluation waves t_0–t_3 to avoid ceiling effects due to children's development in academic skills with age.

Letter Discrimination Task: Our letter discrimination task ("bp task") measures attentional stamina and is taken from Esser *et al.* [13]. In this task, the child saw three lines filled with the letters "b", "d", "g", "q", "h", and "p", in total 45 letters on the touchscreen (see Fig. S13 for an example of such a screen). The child had to go through the letters from left to right, row by row, and tap on all "b"s and "p"s without accidentally marking any other letter. The two target letters "b" and "p" were displayed at the top of the screen in a salient form so that the child was always reminded of the goal in this task in every single trial. The screen emptied after 30 seconds, and a new screen appeared. This was repeated for 18 times (only 12 times in t_0). To construct the outcome score we add up standardized scores for both types of errors (i.e., marking a wrong letter and failure to mark a "b" or a "p"). This score is then again standardized to mean = 0 and SD = 1 within each wave and multiplied by -1. Thus, a higher score indicates better performance in the task (i.e., fewer mistakes).



Example for easy item:

Example for difficult item:

Figure S12: Geometry Task, Screenshot Plus Two Further Examples

1.5 Supplementary Details on the Data Analysis

Details on Descriptive Statistics

Table S1 presents descriptive statistics for the whole sample. Overall, 17 classes (315 children, i.e., 55%) were assigned to the treatment group and 14 classes (257 children) to the control group. About 49% of the children were male, mean age at the beginning of the year (i.e., on January 1, 2013) was 82 months (6.84 years, SD = 0.36 years). Gender and age variables are taken from parental consent forms and are therefore available for all children. The variables migration background and language problems stem from the teacher questionnaire administered in t_0, the variables household income and mothers' university degree stem from the parental questionnaire in t_0. The information about secondary school track choice is taken from a separate parental survey administered three years after the intervention (see Section 1.3).

Estimating the Treatment Effect

To estimate the treatment effect of the self-regulation teaching, we regress outcome scores measured after the treatment on a dummy variable that takes on the value 1 for observations from the treatment group and zero for observations from the control group. To test for treatment effects, we always use two-sided t-tests against the null hypothesis that the coefficient for the treatment dummy is equal to zero. We also control for other important variables in these regressions. The reason for this is as follows.

In the next subsection on "Sample Balance", we will show that there is no evidence for significant imbalances between treatment and control group. However, the absence of evidence for significant imbalances does not mean that control and treatment are perfectly balanced.

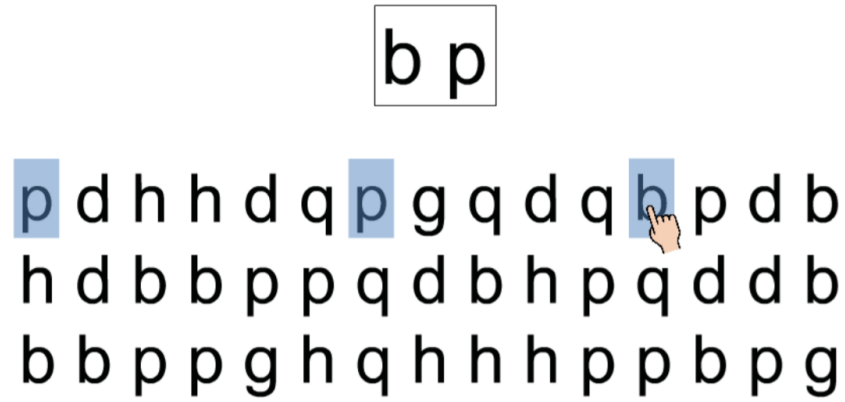


Figure S13: Example of a Screen in the Letter Discrimination Task

With finite samples, there is always the possibility of some imbalance between treatment and the control group in one or another characteristic. It therefore makes sense to control for key demographic characteristics (gender, age, and age at test days) as well as school fixed effects (because randomization was conducted within schools). These controls remove the impact of possible imbalances in these variables and increase the precision of our estimates for the treatment effect. Furthermore, we conducted other treatments (unrelated to the self-regulation teaching) in the same sample, with a randomly chosen part of the self-regulation treatment group and a randomly chosen part of the control group, a working memory training and a learning software training [14]. Of the $n = 315$ children in the self-regulation treatment group, $n = 145$ also received working memory training, $n = 37$ also received learning software training, and $n = 133$ received only self-regulation training. Of the $n = 257$ children in the control group, $n = 134$ received working memory training, $n = 22$ received learning software training, and $n = 101$ received only regular classroom teaching. The other treatments are unrelated to the self-regulation teaching. We control for the other treatments as well as for the interaction effects between the self-regulation treatment and the other treatments in all our estimates of the treatment effect. Finally, we also control for the pre-treatment baseline level of X to estimate the treatment effect of the self-regulation teaching unit on outcomes. Thus, instead of the "difference-in-difference estimator"—which estimates the treatment effect by computing the difference between post-treatment outcome and baseline and then comparing the differences for the control and the treatment group—we control for the baseline level of the respective outcome X. Our treatment estimate thus informs us about how the self-regulation teaching affects outcome X for individuals that have the same given baseline level of X.

The justification for this follows from the fact that "our" estimate of the treatment effect (by controlling for baseline scores) has approximately a variance of

$$\frac{2\sigma^2(1-\rho^2)}{n} \tag{1}$$

while the difference-in-differences estimator has a variance of

$$\frac{4\sigma^2(1-\rho)}{n} \tag{2}$$

where ρ is the autocorrelation of the outcome measures and n the number of observations. The advantage of this method is that the variance of the estimated effect is smaller, i.e., the treatment effect is measured with more precision if $\rho < 1$ [15, 16].

Details on Sample Balance

We provide results on sample imbalance checks in Tables S2 and S3. In our first check (see Table S2), we regress various socio-demographic characteristics such as gender, age, migration

background, etc. on a dummy variable indicating the self-regulation treatment. The table shows that the treatment coefficient in all regressions is close to zero and insignificant, except for one case (language problems rated by the teacher, -0.15 SD, p = .089). Given that we test for eight different socio-demographic variables, it is not surprising that we find a variable with a difference on a 10%-level. Otherwise, there were no significant imbalances between treatment and control group with respect to socio-demographic variables.

As a second sample balance check (see Table S3), we regress standardized baseline test scores (i.e., test scores measured prior to the treatment in t_0) on the treatment indicator, school fixed effects, and the same control variables that are included in the main estimations of the treatment effect (see previous subsection). Again, all coefficients for computer-based test outcomes are insignificant, indicating that there is no evidence for imbalances between treatment and control group. For overall self-regulation abilities, we report a significant difference between treatment and control group. Yet, we still argue that this does not significantly threaten our sample balance, because if we take into account the multiple tests we conduct to ensure that randomization was successful (in total, 14 imbalance test regressions), and correct for this in order to control the family-wise error rate, the effect turns insignificant (p = .162). Also, note that we control for the pre-treatment (i.e., baseline) score of overall self-regulation when we estimate the treatment effect of the self-regulation teaching on overall self-regulation (see Table S5).

Adjusting p-Values for Multiple Testing and More Conservative Clustering

We estimate the treatment effect on several outcome variables at several points in time, i.e., we have a relatively large number of hypotheses. This boosts the probability of wrongly rejecting null hypotheses. If we keep the significance level at 5% for each null hypothesis we test, this implies that the probability of wrongly rejecting each null hypothesis (i.e., detecting a "significant" effect even if there is none) is 5%. However, the probability of rejecting at least one out of many null hypotheses is much larger than 5%. Thus, the probability of over-rejecting (rejecting null hypotheses that should not be rejected, i.e., finding a significant effect when there is none) increases with the number of hypotheses we test. This has to be corrected in order not to arrive at wrong conclusions.

It is worth emphasizing, however, that the time patterns of our results do not suggest that randomly significant findings play a role in our study. The pattern of our results is consistent in the sense that we find an increasing impact of the self-regulation teaching over time on all those variables for which we finally find a significant treatment effect; and furthermore, we find a clear null result across all evaluation waves in those cases in which the treatment had no impact (i.e., math abilities and the letter discrimination task). If the observed significant effects were simply due to randomness and did not reflect true treatment effects, one would expect a more irregular pattern. Nevertheless, it makes sense to check the robustness of our findings with respect to multiple hypothesis testing.

We refrain from using correction methods like those of Bonferroni or Holm, as they do not account for the dependence structure of the underlying data and, thus, lack power. In contrast, we apply the Romano-Wolf stepdown procedure that accounts for the underlying dependence structure to control the family-wise error rate (FWER, see Romano & Wolf [17]—a technique which is increasingly used for large-scale intervention studies (see, for example, [see, for example, 18–20]. Furthermore, we use a newly introduced efficient method to adjust p-values according to this stepdown algorithm [21]. Finally, we also combine this method of controlling the family-wise error rate (FWER) with the BRL (biased-reduced linearization) correction method [22]. This method accounts for potential biases in estimation of standard errors when the number of clusters is relatively small.

We generate families of outcome measures by bundling outcomes in a natural way for the purpose of multiple hypothesis testing. Specifically, we apply the multiple testing correction to control the FWER to all outcomes at all points in time measuring self-regulation teaching-

related outcomes (see Table S4), all outcomes at all points in time measuring broader self-regulation abilities (see Table S5), and outcomes at all points in time measuring non-trained abilities (see Table S6).

We ran $M = 5000$ bootstrap repetitions (stratifying on class-level and correcting standard errors using biased-reduced linearization). Subsequently, we apply the code to adjust p-values according to the stepdown procedure of Romano & Wolf [17, 21]. The resulting p-values are reported in Tables S4–S6 with the label "p-values MHT-BRL". All our main results are robust to using these correction methods.

Details on the Heterogeneity Analysis on Teacher Experience

More experienced teachers are on average better teachers [23]. Therefore, we examine whether our treatment effects are influenced by teacher experience. We create a dummy variable 'Below median Teacher Experience' (used in Table S8) for below-median experienced teachers, based on teachers' years of experience working as a teacher (median teacher experience is 8 years). We interact this dummy with the treatment dummy and add both variables to our main models.

Details on Robustness Checks

In addition to the more conservative estimation methods described above, we estimate tobit models because some of our outcome variables are right-censored (see Fig. S14). We exactly replicate our main models but use tobit instead of least square models. Results are provided in Table S9–S10; all our findings are replicated.

Also, we further analyze the—in general very low—attrition over time in our sample. Attrition across treatment and control group is reported in Table S11. Generally, there are no differences in the probability to drop-out between treatment and control group. A probit model with a dummy for remaining in the sample until t_3 reveals no significant relationship to the treatment dummy (p = .408, controlling for school fixed effects and further controls, see Section 1.5). To further support that children that dropped out of the sample were not driving our findings, we restrict the sample to children who appear in all three follow-up waves. Note that this yields somewhat lower numbers of observation compared to the main results tables because some children appear in t_2 but not in t_3 and vice versa. Results of these estimations can be found in Tables S12–S14; all our findings are robust to this sample restriction.

1.6 Other Supplementary Details

Details on Benchmarking

Here we provide some background information for the benchmarks on treatment effect size discussed in the paper. We evaluate improvements caused by the self-regulation teaching by looking at the distribution of reading abilities in the control group. The raw treatment effect (i.e., prior to standardization of outcomes to mean = 0, SD = 1) amounts to 1.13 raw points in the reading comprehension task in t_3. Applying this improvement to the median child in the control group (median = 13 points) shifts this child from the 50th to the 75th percentile in the distribution.[1] In a similar vein, for the careless mistakes outcome the treatment effect amounts to 1.23 raw points. Adding this to the median score for control group children moves the median control group child from the 50th to the 75th percentile. Taken together, these improvement suggest that the treatment effect is substantial.

To benchmark the reported effect size on secondary school track choice, we compare the 13.3 percentage point increase in the likelihood of choosing an advanced track secondary

[1] To check for robustness, we can also estimate the improvements for children in the 25th and the 75th percentile: a child in the 25th percentile would move up to the 38th percentile, a child in the 75th percentile would move up to the 92nd percentile.

school with the gap in enrollment based on socio-economic status (SES) while controlling for the baseline IQ of the children. More specifically, children in our control group whose mother has a university degree have a 21.4 percentage points higher probability of enrolling into the advanced track compared to children with mothers without a university degree even if we control for baseline IQ (measured in terms of children's fluid IQ). In other words, equally intelligent children have a 21.4 percentage points lower likelihood of enrolling into the advanced track if their mother lacks a university degree. Thus, the effect size of our self-regulation teaching unit on children's enrollment into an advanced track secondary school is equivalent to roughly 62% of the gap that is associated with mothers' university degree.

Details on the Cost-Benefit-Analysis

In order to obtain a (rough) estimate of the costs and benefits of the self-regulation teaching unit, we compute the costs and benefits per child. The costs for the self- regulation teaching unit are calculated by adding up

- Material costs of (a maximum of) €10 per child;
- Costs for the three-hour-training workshop for teachers to teach them the MCII self-regulation strategy and how they can implement it via five teaching lessons:
  - We assume 10 teachers with an average class-size of 20 children per workshop, and an hourly salary for teachers of approx. €26.[2] Therefore, costs for teacher time is roughly equal to €4 per child.
  - In addition, we conservatively estimate fixed costs for a trainer of about €800 per workshop and organizational fixed costs for room, catering, etc. of €400 per workshop. With 10 teachers or 200 children per workshop, there is an additional cost of €6 per child.
- Costs for teachers preparing the teaching lessons because the self-regulation teaching constitutes new material which they have not used before: If teachers prepare each teaching lesson for one full hour (€26 per hour), the preparation cost for five lessons are equal to €130 per class, i.e., about €7 per child.
- Fixed costs for adjusting existing schedules and interrupting regular teaching routines: Every new module in school will come about with some planning and organizational cost. Although this will likely occur at the school- rather than the class-level, we estimate 3 hours of adjustment or planning time per class, i.e., €78 per class or about €4 per child.

The total costs of the self-regulation teaching unit per child is thus given by €10 (material) + €10 (workshop) + €7 (preparation) + €4 (adjustment) = €31 per child (about US$37). We deliberately do not add opportunity cost of time for the teaching lessons for children or teachers because our experimental design tested these lessons against regular classroom teaching, i.e., children and teachers are in school for this period of time in any case.

The benefits of the self-regulation teaching are much harder to quantify. We use two different approaches, a short-term and a long-term perspective. Both approaches are very conservative in estimating the benefits because they are exclusively based on the improvements in reading abilities caused by the self-regulation teaching and, thus, neglect all other improvements in abilities such as inhibition control, overall self-regulation, or the detection and removal of mistakes. Of course, in addition to these benefits one may want to include other non-monetary benefits of improved skills or education in general. The idea of this analysis is

[2]Hourly costs for the teachers are estimated as follows. Depending on experience, teachers in primary school in Germany earn an average monthly gross salary of about €4'500. With 40h/week of working time, teachers earn approx. €26/h.

to generate a lower bound of the value of benefits instead of yielding a precise estimate of the full benefits of the self-regulation teaching unit.

In the first approach, we look at the benefits of the self-regulation teaching on reading abilities and estimate the time (and cost) of schooling necessary to yield these improvements. To gauge the magnitude of our effect size of 0.39 standard deviations after one year, we relate our improvements in reading abilities to the regular development in school in this period of time. The results from a standardized German reading comprehension test (Subtest "Satzverständnistest" from "ELFE II") for our age group reveal the following learning curves for our age group [24, p. 88ff.]: The published norm charts for this test for understanding sentences suggest an effect size of 1.43 standard deviations per year for the time span between t_1 and t_3 (12 months) or about 0.64 SD from t_2 to t_3 (6 months). Hence, the treatment effect of 0.39 SD in reading compares to 3.3 or 3.7 months of reading development (comparing to 12 or 6 months development norm, respectively). German primary schools spend about €6000 per child per year [25]. Of course, total spending is not equal to the investment into reading comprehension and our estimates of the reading improvements are noisy (i.e., have a standard error); therefore, we try to estimate a lower bound by assuming (i) a reading improvement that is equivalent to only one month of the development norm, and (ii) that only 10% of the spending per child per year can be attributed to investments into reading comprehension (most certainly an underestimation given the importance of reading in this age group). Thus, we estimate a benefit of 6000/12 = €500 per month, multiplied by 0.1 = €50 (or about US$60) per child.

In the long run, however, improved reading abilities potentially have much larger (financial and non-financial) returns than the above approach suggests. Therefore, our second approach looks at the existing evidence in the literature how improvements in skills affect lifetime earnings. Hanushek *et al.* [26] use PIACC data and estimate how increases in skill levels affect income. PIACC is the OECD's comprehensive survey of adult literacy and numeracy skills using representative samples in the age group of 16–65 years. They estimate that a one standard deviation increase in PIACC literacy skills is associated with a roughly 17% higher gross hourly wage (pooled across all countries — for Germany, the effect is even larger with 23%, see Hanushek *et al.* [26, Table A3]). Of course, literacy skills in PIAAC data comprise a larger set of skills than our test on reading comprehension. Thus, we conservatively assume that our 0.39 SD increase in reading abilities in our data is comparable to 0.039 SD increase in literacy skills as measured in the PIAAC data. Building on Hampf *et al.* [27], Hanushek & Woessmann [28] use these PIAAC results to estimate effects on lifetime income due to learning losses during the pandemic. We follow their approach and, based on the effect of literacy skills pooled across all countries in Hanushek *et al.* [26], estimate an increase in lifetime income of 0.67% caused by improvements in reading abilities from the self-regulation teaching. Using a conservative estimate of lifetime income of €1'000'000 for Germany [cf. 29], this translates into €6670 higher lifetime earnings over the lifecycle (about US$7950).

In order to calculate the cost-benefit-ratio based on the above discussed approaches, we assume that the the costs (€31) occur at the beginning of the period and that returns arise at the end of the period: for the benefit of €50 after one year; for the benefit of €6670 after 60 years (i.e., conservatively assuming that increased lifetime earnings occur only at the end of the working period). Assuming a discount rate of 5%, we yield cost-to-benefit ratios of 1:1.5 for the first approach and 1:11.5 for the second approach.

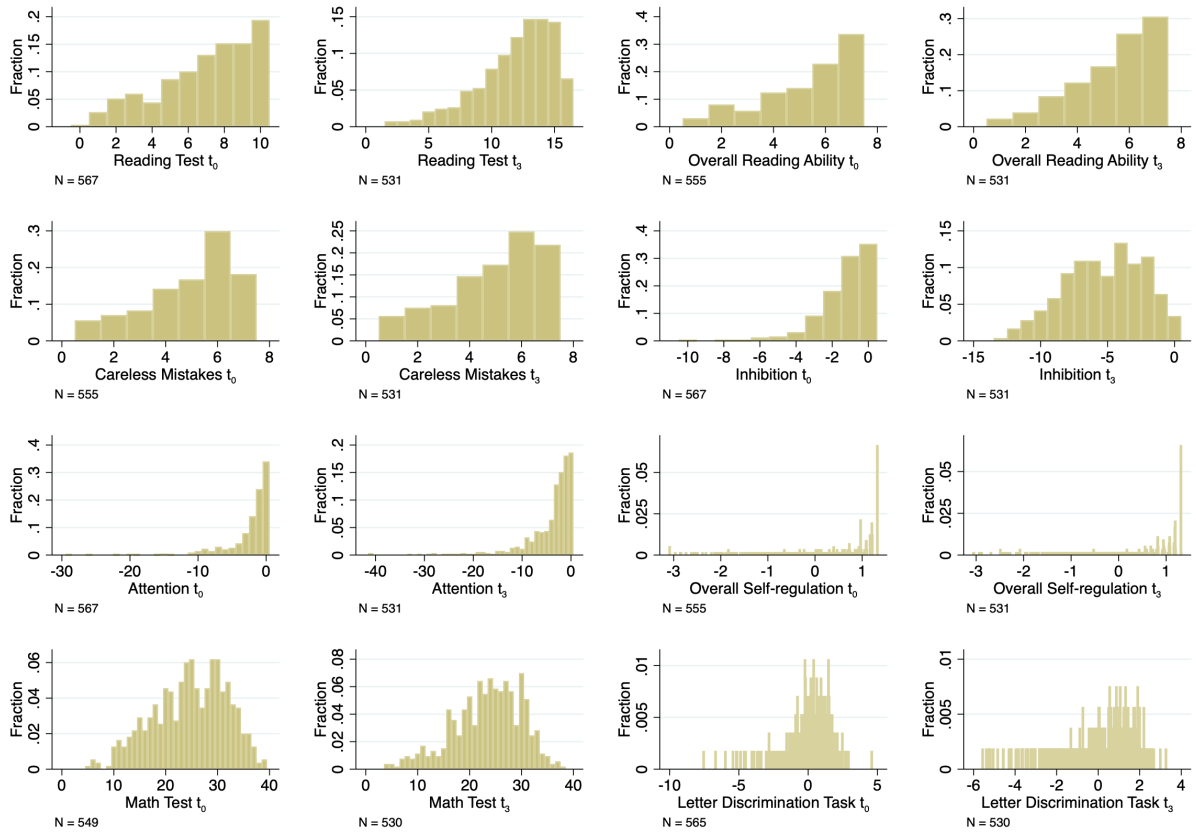
Data Availability Statement

The data for this publication have been collected in a project that has compiled a large set (and combination) of children's abilities, preferences, and family (socio-demographic) characteristics (see Sections 1.3 and 1.4), and thus represents highly sensitive data. This dataset cannot be made available for data protection reasons. In addition, parental consent for data usage only covers strictly scientific purposes. The restriction to scientific purposes was also necessary to comply with data protection requirements and use of the data for strictly scien-

tific purposes cannot be guaranteed if the dataset is made (publicly) available. Not all the data collected in this project are analyzed for this publication, see Section 1.4 for details. Researchers interested in replicating our findings can get access to the data set after filling out a research agreement with us. We confirm that in the paper and the Supplementary Information, we have reported all measures, conditions, data exclusions, and how we determined our sample sizes.

2 Further Supplementary Figures

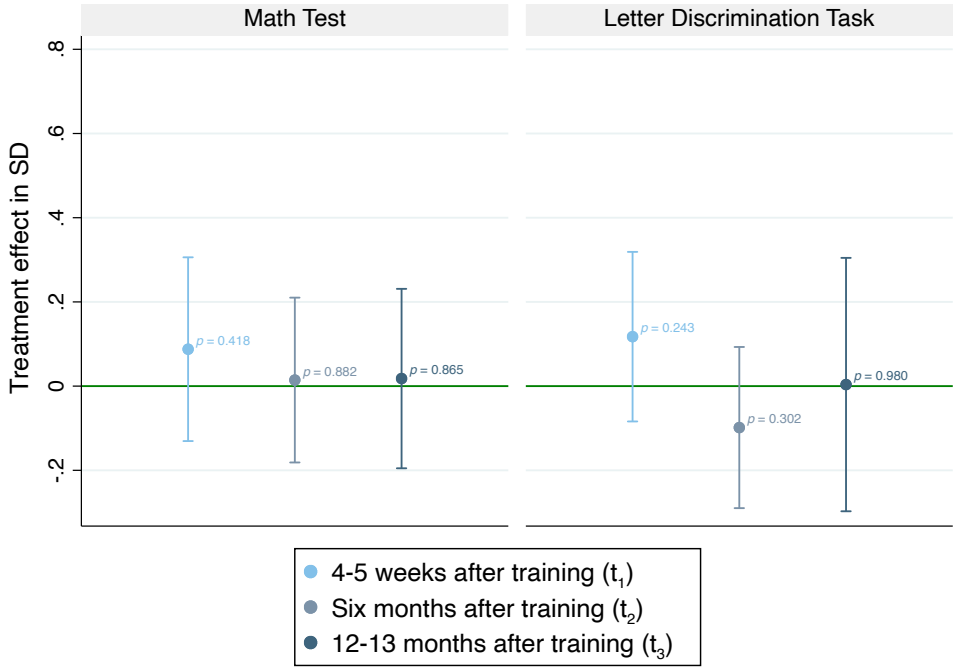
2.1 Distribution of Main Outcome Scores



Notes: The figure shows the distribution of raw outcome scores (i.e, before standardization) at baseline (t_0) and 12–13 months after the self-regulation teaching (t_3). Note that scores are not comparable across waves because the difficulty for each test had to be adjusted over time to account for the development of skills (see Section 1.5 for details).

Figure S14: Distribution of Nonstandardized t_0 and t_3 Outcome Scores

2.2 Treatment Effects on Other Academic Domains



Notes: The dots show point estimates (as fractions of a standard deviation of the respective outcome) of how self-regulation teaching changes the outcome relative to the control group in domains (indicated by the panel title) that were not directly targeted by the self-regulation teaching. *Math Test* is the aggregate test score in arithmetic and geometry tasks and *Letter Discrimination Task* measures performance in the bp task. The bars indicate 95% confidence intervals. All estimates are based on least squares models controlling for school fixed effects, pre-treatment outcome scores, and further controls (more details and p-values adjusted for multiple hypothesis testing are reported in Table S6). Standard errors are clustered at the classroom level.

Figure S15: Treatment Effects on Math Test and Letter Discrimination Task

3 Supplementary Tables

3.1 Summary Statistics

Table S1: Descriptive Statistics

	Mean	SD	Min	Max	N
Self-regulation Teaching	0.55	0.50	0	1	572
Male	0.49	0.50	0	1	572
Children's age in months on Jan 1, 2013	82.13	4.32	72	102	572
Children's age on test day W1 (in months)	84.25	4.38	75	103	572
Children's age on test day W2 (in months)	87.29	4.35	78	107	572
Children's age on test day W3 (in months)	92.37	4.38	83	112	544
Children's age on test day W4 (in months)	99.58	4.38	90	119	531
Migration background	0.45	0.50	0	1	568
Language problems	0.25	0.43	0	1	572
Monthly HH-Net Income <1500 EUR	0.14	0.35	0	1	441
Monthly HH-Net Income 1500–2500 EUR	0.21	0.41	0	1	441
Monthly HH-Net Income 2500–5000 EUR	0.43	0.50	0	1	441
Monthly HH-Net Income >5000 EUR	0.22	0.41	0	1	441
Mother university degree	0.45	0.50	0	1	444
Mother vocational degree	0.42	0.49	0	1	444
Mother no professional degree	0.13	0.34	0	1	444
Academic track secondary school	0.69	0.46	0	1	393
Mixed-track secondary school	0.20	0.40	0	1	393
Non-academic track secondary school	0.10	0.31	0	1	393

Notes: The table provides socio-demographic information about our sample. Gender and age information was reported on the parental consent form and is therefore available for all children. The variables 'migration background' and 'language problems' are taken from the teacher questionnaire in t_0; for four children teachers reported not to know the migration background. Income and maternal education variables are taken from the parent questionnaire in t_0. The information about secondary school track choice is taken from a survey administered to parents three years after the treatment (see Section 1.3 for details).

3.2 Sample Balance

Table S2: Sample Balance for Socio-demographic Variables

	(1) Male	(2) Age in months	(3) Migration Backgr.	(4) Language Probl.	(5) HH Income <1500	(6) HH Income >2500	(7) Mother no degree	(8) Mother University
Self-regulation	0.071	-0.826	-0.150	-0.154*	-0.044	0.076	0.013	-0.075
Teaching	(0.052)	(0.576)	(0.121)	(0.082)	(0.071)	(0.091)	(0.062)	(0.096)
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Further controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	572	572	568	572	441	441	444	444

Notes: The results are based on least squares models including school fixed effects and further controls (except for demographic variables, see Section 1.5). Standard errors in parentheses are clustered at the school level. * p<.10, ** p<.05, *** p<.01. The coefficients for 'Self-regulation Teaching' in the first row and the associated p-values indicate whether there are significant imbalances between the treatment and control group with respect to the socio-demographic characteristics described in the column titles. Since it turns out that all coefficients are close to zero or small and no coefficient is significantly different from zero at the 5% level, there is no evidence for significant imbalances (the difference for language problems on a 10%-level does not go beyond random chance when testing for eight different sociodemographic variables). The sample size in column 3 is smaller than the total sample size because the dependent variable 'migration background' is taken from the teacher questionnaire and for four children teachers reported not to know the migration background. Sample sizes in columns 5–8 are smaller because the dependent variables are taken from the parent questionnaire which was not answered by all parents.

Table S3: Sample Balance for Outcomes at Baseline (t_0)

	(1) Reading Test	(2) Ov. Reading Ability	(3) Careless Mistakes	(4) Inhibition	(5) Attention	(6) Overall Self-regulation	(7) Math Test	(8) Letter Disc. Task
Self-regulation	-0.134	0.024	0.016	-0.177	-0.048	0.372***	-0.073	-0.226
Teaching	(0.282)	(0.288)	(0.270)	(0.148)	(0.121)	(0.094)	(0.153)	(0.141)
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Further controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	567	555	555	567	567	555	549	565

Notes: The results are based on least squares models including school fixed effects and further controls (see Section 1.5). All outcome scores are standardized to mean $=$ 0 and SD $=$ 1. Standard errors in parentheses are clustered at the school level. * p<.10, ** p<.05, *** p<.01. The coefficients in the first row and the associated p-values indicate whether there are significant imbalances between the treatment and control group regarding the respective baseline outcome measures. It turns out that all coefficients for 'Self-regulation Teaching' (except the one for overall self-regulation) are close to zero and insignificant at the 10% level, i.e., there is no evidence for significant imbalances between treatment and control group for these outcome measures. Given that we test for eight different baseline scores, finding one outcome with significant differences between treatment and control does not indicate that randomization as a whole would have failed. In addition, we always control for the baseline score of overall self-regulation when we estimate the treatment effect of self-regulation teaching on overall self-regulation (see Table S5; for a justification of this control see Section 1.5).

3.3 Main Results

Table S4: Treatment Effects on Self-regulation Teaching Outcomes

	Reading Test			Overall Reading Ability			Careless Mistakes Measure		
	t_1	t_2	t_3	t_1	t_2	t_3	t_1	t_2	t_3
Self-regulation Teaching	0.203**	0.205	0.391***	0.002	0.288**	0.366***	0.025	0.474**	0.691***
	(0.082)	(0.125)	(0.132)	(0.091)	(0.140)	(0.119)	(0.136)	(0.179)	(0.177)
	[0.019]	[0.111]	[0.006]	[0.983]	[0.049]	[0.005]	[0.858]	[0.013]	[0.001]
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Further Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
p-values MHT-BRL	0.063*	0.167	0.014**	0.974	0.063*	0.010***	0.963	0.024**	0.002***
N	564	539	526	555	527	517	555	527	517

Notes: The results are based on least squares models that regress our self-regulation teaching outcome measures on a dummy variable that takes on the value of 1 if the child received 'Self-regulation Teaching' and 0 otherwise. The regression also includes school fixed effects, the baseline outcome score (t_0), and further controls (see Section 1.5). All outcome scores are standardized to mean $=$ 0 and SD $=$ 1. t_1, t_2, and t_3 refer to the evaluation waves shortly after, 6 months after, and 12–13 months after the self-regulation teaching. The coefficients in the first row are the point estimates showing how self-regulation teaching changes the outcome score indicated at the top of the table (as a fraction of a standard deviation) relative to the control group. Standard errors in parentheses are clustered at the classroom level. Related p-values are reported in brackets (* p<.10, ** p<.05, *** p<.01). We conduct a further robustness check by adjusting p-values for multiple hypothesis testing (MHT) by controlling the family-wise error rate using the step-down procedure by Romano & Wolf [17, 21], while at the same time applying the more conservative "biased reduced linearization (BRL) method" of Bell & McCaffrey [22] to calculate clustered standard errors. Based on this more conservative estimation method, the treatment effects on reading abilities in t_3 remains significant on the 5% and the 1% level, respectively, and the treatment effect on careless mistakes in t_3 remains significant on the 1% level.

Table S5: Treatment Effects on Self-regulation Abilities

	Inhibition			Attention			Overall Self-Regulation		
	t_1	t_2	t_3	t_1	t_2	t_3	t_1	t_2	t_3
Self-regulation Teaching	0.177	-0.045	0.261***	0.205	0.127	0.559***	0.299***	0.291**	0.568***
	(0.118)	(0.103)	(0.065)	(0.150)	(0.110)	(0.118)	(0.072)	(0.106)	(0.082)
	[0.146]	[0.667]	[0.000]	[0.181]	[0.258]	[0.000]	[0.000]	[0.010]	[0.000]
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Further Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
p-values MHT-BRL	0.422	0.667	0.005***	0.422	0.459	0.005***	0.005***	0.066*	0.001***
N	566	540	527	565	540	527	555	527	517

Notes: The results are based on least squares models that regress the various self-regulation ability measures on a dummy variable that takes on the value of 1 if the child received 'Self-regulation Teaching' and 0 otherwise. The regression also includes school fixed effects, the baseline outcome score (t_0), and further controls (see Section 1.5). All outcome scores are standardized to mean $=$ 0 and SD $=$ 1. t_1, t_2, and t_3 refer to the evaluation waves shortly after, 6 months after, and 12–13 months after the self-regulation teaching. The coefficients in the first row are the point estimates showing how self-regulation teaching changes the outcome score indicated at the top of the table (as a fraction of a standard deviation) relative to the control group. Standard errors in parentheses are clustered at the classroom level. Related p-values are reported in brackets (* p<.10, ** p<.05, *** p<.01). We conduct a further robustness check by adjusting p-values for multiple hypothesis testing (MHT) by controlling the family-wise error rate using the step-down procedure by Romano & Wolf [17, 21], while at the same time applying the more conservative "biased reduced linearization (BRL) method" of Bell & McCaffrey [22] to calculate clustered standard errors. Based on this more conservative estimation method, the treatment effects on inhibition and attention in t_3 remain significant on the 1% level. Similarly, the treatment effects on overall self-regulation abilities in t_1 and t_3 remain significant on the 1% level.

Table S6: Treatment Effects on Other Academic Domains

	Math Test			**Letter Discrimination Task**		
	t_1	t_2	t_3	t_1	t_2	t_3
Self-regulation Teaching	0.088	0.014	0.018	0.117	-0.098	0.004
	(0.107)	(0.096)	(0.104)	(0.099)	(0.094)	(0.147)
	[0.418]	[0.882]	[0.865]	[0.243]	[0.302]	[0.980]
School FE	Yes	Yes	Yes	Yes	Yes	Yes
Further Controls	Yes	Yes	Yes	Yes	Yes	Yes
p-values MHT-BRL	0.797	0.995	0.995	0.736	0.764	0.995
N	535	525	511	552	538	524

Notes: The results are based on least squares models that regress our academic transfer outcomes on a dummy variable that takes on the value of 1 if the child received 'Self-regulation Teaching' and 0 otherwise. The regression also includes school fixed effects, the baseline outcome score (t_0), and further controls (see Section 1.5). All outcome scores are standardized to mean $=$ 0 and SD $=$ 1. t_1, t_2, and t_3 refer to the evaluation waves shortly after, 6 months after, and 12–13 months after the self-regulation teaching. The coefficients in the first row are the point estimates showing how self-regulation teaching changes the outcome score indicated at the top of the table (as a fraction of a standard deviation) relative to the control group. Standard errors in parentheses are clustered at the classroom level. Related p-values are reported in brackets (* p<.10, ** p<.05, *** p<.01). We conduct a further robustness check by adjusting p-values for multiple hypothesis testing (MHT) by controlling the family-wise error rate using the step-down procedure by Romano & Wolf [17, 21], while at the same time applying the more conservative "biased reduced linearization (BRL) method" of Bell & McCaffrey [22] to calculate clustered standard errors.

Table S7: Treatment Effects on Secondary School Track Choice Three Years After Treatment

	Advanced Track (OLS)		**Advanced Track (IPW)**	
	(1)	(2)	(3)	(4)
Self-regulation Teaching	0.133***	-0.044	0.158***	-0.034
	(0.045)	(0.054)	(0.043)	(0.054)
	[0.006]	[0.422]	[0.001]	[0.532]
Reading Test t_3		0.153***		0.168***
		(0.028)		(0.028)
Careless Mistakes t_3		0.063**		0.060**
		(0.030)		(0.029)
Inhibition t_3		0.004		0.002
		(0.024)		(0.023)
Attention t_3		0.007		-0.000
		(0.026)		(0.025)
Overall Self-regulation t_3		0.083***		0.081***
		(0.025)		(0.025)
School FE	Yes	Yes	Yes	Yes
Further Controls	Yes	Yes	Yes	Yes
Inverse Probability Weighting	No	No	Yes	Yes
p-value BRL	0.036**		0.024**	
N	393	393	393	393

Notes: Columns (1) and (2) show regression results that are based on ordinary least square models (OLS). To check the robustness of our results to attrition, we report regressions based on inverse probability weighting (IPW) in columns (3) and (4). In all columns, we control for school fixed effects and further controls (see Section 1.5). The dependent variable in all columns is a binary variable taking on the value of 1 if the child is enrolled in an advanced track secondary school three years after treatment and the value of 0 if the child is enrolled in another secondary school. Standard errors in parentheses are clustered at the classroom level. * p<.10, ** p<.05, *** p<.01. Column (1) demonstrates that children in the treatment group have a 13-percentage point higher likelihood of enrolling in an advanced track secondary school (p = .006). We conduct a further robustness check by adjusting p-values applying the more conservative "biased reduced linearization (BRL) method" of Bell & McCaffrey [22] to calculate clustered standard errors. Based on this more conservative estimate, the treatment effect on choosing an advanced track secondary school three years after treatment remains significant on the 5% level. In column (2), we provide evidence supporting that this treatment effect is mediated by treatment-induced improvements in reading abilities, overall self-regulation abilities, and better skills in finding careless mistakes in t_3 (i.e., treatment outcomes measured 12 months after the treatment). All mediator variables are standardized to mean $=$ 0 and SD $=$ 1. Adding improved self-regulation teaching outcomes and self-regulation abilities in t_3 diminishes the effect of the treatment on choosing an advanced track secondary school to about zero (see column (2). Thus, our data support that the treatment-induced improvements in reading, overall self-regulation, and the finding of careless mistakes are a plausible mechanism that is underlying the impact of the self-regulation teaching on school track choice. Columns (3) and (4) provide a robustness check accounting for attrition. We use inverse probability weighting with weights based on three binary variables, migration background, baseline educational achievement (median split), and baseline cognitive skills (median split). Educational achievement is constructed using the sum of standardized scores in math and reading in t_0. Cognitive skills are calculated using the sum of standardized scores in working memory capacity and Raven's fluid IQ. To calculate the weights, missing values are imputed with the sample mean. Results confirm that attrition does not seem to be an important factor for the size and significance of the treatment effect.

3.4 Heterogeneity Analysis w.r.t. Teacher Experience

Table S8: Heterogeneous Treatment Effects by Teacher Experience

	Reading Test	**Ov. Reading Ability**	**Careless Mistakes**	**Inhibition**	**Attention**	**Ov. Self-regulation**
	t_3	t_3	t_3	t_3	t_3	t_3
Self-regulation	0.422***	0.388***	0.734***	0.305***	0.563***	0.584***
Teaching	(0.147)	(0.121)	(0.178)	(0.066)	(0.116)	(0.096)
SRT x Below median	-0.149	-0.195	0.448	-0.205	0.029	-0.197
Teacher Exp	(0.194)	(0.249)	(0.422)	(0.123)	(0.171)	(0.208)
School FE	Yes	Yes	Yes	Yes	Yes	Yes
Further Controls	Yes	Yes	Yes	Yes	Yes	Yes
N	526	517	517	527	527	517

Notes: This table examines whether the treatment effects of 'Self-regulation Teaching' differ for teachers that are more or less experienced (as a proxy for teaching quality). For this purpose, we include a dummy variable 'Below-median Teacher Experience' which takes on a value of 1 if the teacher show a below-median number of years of professional experience at baseline and a value of 0 otherwise. We then interact 'Below median Teacher Experience' with the treatment dummy. In addition, the least squares regressions also include school fixed effects, the baseline outcome score (t_0), and further controls (see Section 1.5). All outcome scores are standardized to mean $=$ 0 and SD $=$ 1. t_3 refers to the evaluation wave 12–13 months after the self-regulation teaching. Standard errors in parentheses are clustered at the classroom level. * p<.10, ** p<.05, *** p<.01. In this table, the coefficient related to the treatment dummy shows the treatment effect for children with a teacher who reports above-median years of experience. The coefficient of the interaction term 'SRT $\times$ Below-median Teacher Experience' shows the extent to which the treatment effect is different for children with teachers having a below-median experience. None of the interaction terms is significantly different from zero, indicating that classes with teachers with a below-median experience do not show a significantly lower treatment effect.

3.5 Tobit Estimates of Treatment Effects

Table S9: Treatment Effects on Self-regulation Teaching Outcomes Using Tobit Models

	Reading Test			**Overall Reading Ability**			**Careless Mistakes Measure**		
	t_1	t_2	t_3	t_1	t_2	t_3	t_1	t_2	t_3
Self-regulation Teaching	0.220**	0.186	0.415***	-0.063	0.373*	0.517***	0.025	0.474***	0.836***
	(0.102)	(0.134)	(0.122)	(0.108)	(0.212)	(0.148)	(0.134)	(0.176)	(0.203)
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Further Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	564	539	526	555	527	517	555	527	517

Notes: The table reports results for the models estimated in Table S4 but using Tobit models in order to take the potential censoring of the dependent variable into account. We regress the various outcome scores on a dummy variable that takes on the value of 1 if the child received 'Self-regulation Teaching' and 0 otherwise. The regression also includes school fixed effects, the baseline outcome score (t_0), and further controls (see Section 1.5). All outcome scores are standardized to mean $= 0$ and SD $= 1$. t_1, t_2, and t_3 refer to the evaluation waves shortly after, 6 months after, and 12–13 months after the self-regulation teaching. Standard errors in parentheses are clustered at the classroom level. * p<.10, ** p<.05, *** p<.01. The coefficients in the first row show the treatment effect; all results are in line with the findings from Table S4.

Table S10: Treatment Effects on Self-regulation Abilities Using Tobit Models

	Inhibition			**Attention**			**Overall Self-Regulation**		
	t_1	t_2	t_3	t_1	t_2	t_3	t_1	t_2	t_3
Self-regulation Teaching	0.191	-0.024	0.282***	0.249	0.147	0.580***	0.298***	0.291***	0.568***
	(0.117)	(0.106)	(0.067)	(0.156)	(0.135)	(0.126)	(0.071)	(0.104)	(0.080)
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Further Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	566	540	527	565	540	527	555	527	517

Notes: The table reports results for the models estimated in Table S5 but using Tobit models in order to take the potential censoring of the dependent variable into account. We regress the various outcome scores on a dummy variable that takes on the value of 1 if the child received 'Self-regulation Teaching' and 0 otherwise. The regression also includes school fixed effects, the baseline outcome score (t_0), and further controls (see Section 1.5). All outcome scores are standardized to mean $= 0$ and SD $= 1$. t_1, t_2, and t_3 refer to the evaluation waves shortly after, 6 months after, and 12–13 months after the self-regulation teaching. Standard errors in parentheses are clustered at the classroom level. * p<.10, ** p<.05, *** p<.01. The coefficients in the first row show the treatment effect; all results are in line with the findings from Table S5.

3.6 Restricting the Analyses to the No-Attrition-Sample

Table S11: Attrition in Treatment and Control Group

Attrition	Control Group	Treatment Group	Total
Remain in sample	235	296	531
Lost from sample	22	19	41
Total	257	315	572

Notes: The table reports the number of children that stay in the sample for all evaluation waves (from t_0–t_3) across treatment and control group.

Table S12: Treatment Effects on Self-regulation Teaching Outcomes—No Attrition

	Reading Test			**Overall Reading Ability**			**Careless Mistakes Measure**		
	t_1	t_2	t_3	t_1	t_2	t_3	t_1	t_2	t_3
Self-regulation Teaching	0.222***	0.210	0.410***	-0.003	0.298**	0.366***	0.016	0.459**	0.691***
	(0.063)	(0.126)	(0.129)	(0.095)	(0.140)	(0.119)	(0.143)	(0.182)	(0.177)
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Further Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	524	524	524	517	517	517	517	517	517

Notes: This table estimates the treatment effects of the self-regulation teaching (see Table S4) for the sample restricted to those children that remain in the sample for all evaluation waves (from t_0–t_3). The results are based on least squares models that regress the various self-regulation ability measures on a dummy variable that takes on the value of 1 if the child received 'Self-regulation Teaching' and 0 otherwise. The regression also includes school fixed effects, the baseline outcome score (t_0), and further controls (see Section 1.5). All outcome scores are standardized to mean $=$ 0 and SD $=$ 1. t_1, t_2, and t_3 refer to the evaluation waves shortly after, 6 months after, and 12–13 months after the self-regulation teaching. The coefficients in the first row are the point estimates showing how self-regulation teaching changes the outcome score indicated at the top of the table (as a fraction of a standard deviation) relative to the control group. Standard errors in parentheses are clustered at the classroom level. * p<.10, ** p<.05, *** p<.01. All findings from Table S4 are replicated.

Table S13: Treatment Effects on Self-regulation Abilities—No Attrition

	Inhibition			**Attention**			**Overall Self-Regulation**		
	t_1	t_2	t_3	t_1	t_2	t_3	t_1	t_2	t_3
Self-regulation Teaching	0.193	0.016	0.260***	0.168	0.128	0.570***	0.299***	0.298**	0.568***
	(0.118)	(0.096)	(0.065)	(0.144)	(0.110)	(0.121)	(0.074)	(0.112)	(0.082)
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Further Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	526	526	526	525	525	525	517	517	517

Notes: This table estimates the treatment effects of the self-regulation teaching (see Table S5) for the sample restricted to those children that remain in the sample for all evaluation waves (from t_0–t_3). The results are based on least squares models that regress the various self-regulation ability measures on a dummy variable that takes on the value of 1 if the child received 'Self-regulation Teaching' and 0 otherwise. The regression also includes school fixed effects, the baseline outcome score (t_0), and further controls (see Section 1.5). All outcome scores are standardized to mean $=$ 0 and SD $=$ 1. t_1, t_2, and t_3 refer to the evaluation waves shortly after, 6 months after, and 12–13 months after the self-regulation teaching. The coefficients in the first row are the point estimates showing how self-regulation teaching changes the outcome score indicated at the top of the table (as a fraction of a standard deviation) relative to the control group. Standard errors in parentheses are clustered at the classroom level. * p<.10, ** p<.05, *** p<.01. All findings from Table S5 are replicated.

Table S14: Treatment Effects on Academic Abilities—No Attrition

	Math Test			**Letter Discrimination Task**		
	t_1	t_2	t_3	t_1	t_2	t_3
Self-regulation Teaching	0.135	-0.013	-0.006	0.138	-0.079	-0.026
	(0.119)	(0.075)	(0.109)	(0.102)	(0.112)	(0.155)
School FE	Yes	Yes	Yes	Yes	Yes	Yes
Further Controls	Yes	Yes	Yes	Yes	Yes	Yes
N	498	498	498	512	512	512

Notes: This table estimates the treatment effects of the self-regulation teaching (see Table S6) for the sample restricted to those children that remain in the sample for all evaluation waves (from t_0–t_3). The results are based on least squares models that regress the various self-regulation ability measures on a dummy variable that takes on the value of 1 if the child received 'Self-regulation Teaching' and 0 otherwise. The regression also includes school fixed effects, the baseline outcome score (t_0), and further controls (see Section 1.5). All outcome scores are standardized to mean $=$ 0 and SD $=$ 1. t_1, t_2, and t_3 refer to the evaluation waves shortly after, 6 months after, and 12–13 months after the self-regulation teaching. The coefficients in the first row are the point estimates showing how self-regulation teaching changes the outcome score indicated at the top of the table (as a fraction of a standard deviation) relative to the control group. Standard errors in parentheses are clustered at the classroom level. * p<.10, ** p<.05, *** p<.01. All findings from Table S6 are replicated.

3.7 Parental Ratings of Self-regulation

Table S15: Treatment Effects on Overall Self-regulation as Rated by Parents

	Overall Self-regulation (Parents)
	(1)
Self-regulation Teaching	0.130
	(0.087)
School FE	Yes
Further Controls	Yes
N	363

Notes: The results are based on least squares models that regress the parent-rated self-regulation measure on a dummy variable that takes on the value of 1 if the child received 'Self-regulation Teaching' and 0 otherwise. The regression also includes school fixed effects, the baseline outcome score (t_0), and further controls (see Section 1.5). All outcome scores are standardized to mean $=$ 0 and SD $=$ 1. t_2 refers to the evaluation waves 6 months after the self-regulation teaching. The coefficients in the first row are the point estimates showing how self-regulation teaching changes the outcome score indicated at the top of the table (as a fraction of a standard deviation) relative to the control group. Standard errors in parentheses are clustered at the classroom level (* p<.10, ** p<.05, *** p<.01).

4 References SI

1. Oettingen, G. Future thought and behaviour change. *European Review of Social Psychology* **23,** 1–63 (2012).
2. Gollwitzer, P. M. Implementation intentions. *American Psychologist* **54,** 493–503 (1999).
3. Gollwitzer, P. M. & Sheeran, P. Implementation intentions and goal achievement: A meta-analysis of effects and processes. *Advances in Experimental Social Psychology* **38,** 69–119 (2006).
4. Gollwitzer, P. M. & Oettingen, G. Goal Pursuit. *The Oxford Handbook of Human Motivation* (2012).
5. Gawrilow, C., Morgenroth, K., Schultz, R., Oettingen, G. & Gollwitzer, P. M. Mental contrasting with implementation intentions enhances self-regulation of goal pursuit in schoolchildren at risk for ADHD. *Motivation and Emotion* **37,** 134–145 (2013).
6. Duckworth, A. L., Kirby, T. A., Gollwitzer, A. & Oettingen, G. From Fantasy to Action Mental Contrasting With Implementation Intentions (MCII) Improves Academic Performance in Children. *Social Psychological and Personality Science* **4,** 745–753 (2013).
7. Gawrilow, C. & Gollwitzer, P. M. Implementation Intentions Facilitate Response Inhibition in Children with ADHD. *Cognitive Therapy and Research* **32,** 261–280 (2008).
8. Goodman, R. The Strengths and Difficulties Questionnaire: A Research Note. *Journal of Child Psychology and Psychiatry* **38,** 581–586 (1997).
9. Tangney, J. P., Baumeister, R. F. & Boone, A. L. High Self-Control Predicts Good Adjustment, Less Pathology, Better Grades, and Interpersonal Success. *Journal of Personality* **72,** 271–324 (2004).
10. Bertrams, A. & Dickhäuser, O. Messung dispositioneller Selbstkontroll-Kapazität. *Diagnostica* **55,** 2–10 (2009).
11. Duckworth, A. L. & Kern, M. L. A meta-analysis of the convergent validity of self-control measures. *Journal of Research in Personality* **45,** 259–268 (2011).
12. Enkavi, A. Z. *et al.* Large-scale analysis of test-retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences* **116,** 5472–5477 (2019).
13. Esser, G., Wyschkon, A. & Ballaschk, K. *Basisdiagnostik Umschriebener Entwicklungsstörungen im Grundschulalter (BUEGA)* (Hogrefe, Göttingen, 2008).
14. Berger, E. M., Fehr, E., Hermes, H., Schunk, D. & Winkel, K. *The Impact of Working Memory Training on Children's Cognitive and Noncognitive Skills* Working Paper (University of Mainz, 2022).
15. McKenzie, D. Beyond baseline and follow-up: The case for more T in experiments. *Journal of Development Economics* **99,** 210–221 (2012).
16. Frison, L. & Pocock, S. J. Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design. *Statistics in Medicine* **11,** 1685–1704 (1992).
17. Romano, J. P. & Wolf, M. Stepwise Multiple Testing as Formalized Data Snooping. *Econometrica* **73,** 1237–1282 (2005).
18. Heckman, J., Moon, S. H., Pinto, R., Savelyev, P. & Yavitz, A. Analyzing social experiments as implemented: A reexamination of the evidence from the HighScope Perry Preschool Program. *Quantitative Economics* **1,** 1–46 (2010).
19. Gertler, P. *et al.* Labor market returns to an early childhood stimulation intervention in Jamaica. *Science* **344,** 998–1001 (2014).

20. Campbell, F. *et al.* Early Childhood Investments Substantially Boost Adult Health. *Science* **343,** 1478–1485 (2014).
21. Romano, J. P. & Wolf, M. Efficient computation of adjusted p-values for resampling-based stepdown multiple testing. *Statistics & Probability Letters* **113,** 38–40 (2016).
22. Bell, R. M. & McCaffrey, D. F. Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples. *Survey Methodology* **28,** 169–181 (2002).
23. Hanushek, E. A. & Rivkin, S. G. in *Handbook of the Economics of Education* 1051–1078 (Elsevier, 2006).
24. Lenhard, W., Lenhard, A. & Schneider, W. *ELFE II - Ein Leseverständnistest für Erst- bis Siebtklässler – Version II (3. Auflage)* tech. rep. (hogrefe Schultests, 2018).
25. Schnitzlein, D. Low Level of Equal Opportunities in Germany: Family Background Shapes Individual Economic Success. *DIW Economic Bulletin* **3,** 3–8 (2013).
26. Hanushek, E. A., Schwerdt, G., Wiederhold, S. & Woessmann, L. Returns to skills around the world: Evidence from PIAAC. *European Economic Review* **73,** 103–130 (2015).
27. Hampf, F., Wiederhold, S. & Woessmann, L. Skills, earnings, and employment: exploring causality in the estimation of returns to skills. *Large-scale Assessments in Education* **5,** 12 (2017).
28. Hanushek, E. A. & Woessmann, L. *The economic impacts of learning losses* OECD Education Working Papers 225 (2020).
29. Stüber, H. *Berufsspezifische Lebensentgelte: Qualifikation zahlt sich aus* IAB-Kurzbericht 17/2016 (IAB, Nürnberg, 2016).