



**University of
Zurich** ^{UZH}

University of Zurich
Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 347

The Impact of Working Memory Training on Children's Cognitive and Noncognitive Skills

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

June 2020

The Impact of Working Memory Training on Children's Cognitive and Noncognitive Skills

Eva M. Berger¹, Ernst Fehr², Henning Hermes³, Daniel Schunk¹, Kirsten Winkel¹

8 June 2020

Abstract: Working memory capacity is thought to play an important role for a wide range of cognitive and noncognitive skills such as fluid intelligence, math, reading, the inhibition of pre-potent impulses or more general self-regulation abilities. Because these abilities substantially affect individuals' life trajectories in terms of health, education, and earnings, the question of whether working memory (WM) training can improve them is of considerable importance. However, whether WM training leads to improvements in these far-transfer skills is contested. Here, we examine the causal impact of WM training embedded in regular school teaching by a randomized educational intervention involving a sample of 6–7 years old first graders. We find substantial immediate and lasting gains in working memory capacity. In addition, we document relatively large positive effects on geometry skills, reading skills, Raven's fluid IQ measure, the ability to inhibit pre-potent impulses and self-regulation abilities. Moreover, these far-transfer effects emerge over time and only become fully visible after 12–13 months. Finally, we document that 3–4 years after the intervention, the children who received training have a roughly 16 percentage points higher probability of entering the academic track in secondary school.

¹Johannes Gutenberg University of Mainz, Department of Law and Economics, Jakob-Welder-Weg 4, 55128 Mainz, Germany. eva.berger@uni-mainz.de, daniel.schunk@uni-mainz.de, kirsten.winkel@uni-mainz.de

²University of Zurich, Department of Economics, Blümlisalpstrasse 10, 8006 Zurich, Switzerland. ernst.fehr@econ.uzh.ch

³Norwegian School of Economics, FAIR / Department of Economics, Helleveien 30, 5045 Bergen, Norway. henning.hermes@nhh.no

I. Introduction

Cognitive and noncognitive¹ skills affect important individual life outcomes such as health, education, and earnings (Cunha et al. 2006; Heckman, Stixrud and Urzua 2006; Moffitt et al. 2011; Duckworth et al. 2012; Almond, Currie and Duque 2018). Working memory (WM) capacity—the ability to mentally store and process information (Baddeley 1999)—is thought to play a key role in a wide range of these abilities. WM capacity is, for example, positively associated with math and language skills (Gathercole et al. 2004; Alloway and Alloway 2010), general fluid IQ (Kyllonen and Christal 1990; Ackerman, Beier and Boyle 2005; Oberauer et al. 2005; de Abreu, Conway and Gathercole 2010) and self-regulation skills such as attention and inhibitory ability (Engle 2002; Hofmann et al. 2008; Schmeichel, Volokhov and Darnree 2008; Diamond and Ling 2020). Conversely, individuals with learning problems, self-regulation and attention deficits often have low WM capacity (Westerberg et al. 2004; Martinussen et al. 2005; Van Snellenberg et al. 2016). In view of the close association between WM capacity and many important skills, the question is whether one can *simultaneously* improve several of these skills through WM training and, if so, whether schools should consider introducing WM training into the curriculum. These questions are of fundamental importance for human capital formation and its underlying mechanisms as well as for educational policy.

Previous evidence suggests that WM training can improve performance on untrained WM tasks (“near-transfer effects”). However, the question of whether training-induced improvements in WM capacity lead to improvements in other important skills, such as academic and self-regulation skills (“far-transfer effects”), lacks a conclusive answer, as even meta-analyses and review studies are controversial on this point (Shipstead, Hicks and Engle 2012; Karbach and Verhaeghen 2014; Au et al. 2015; Melby-Lervag, Redick and Hulme 2016; Aksayli, Sala and Gobet 2019; Sala et al. 2019).

This lack of a conclusive answer suggests that WM-training studies face considerable challenges such as the problem that (i) far-transfer effects are likely to need time to evolve and identifying these effects requires follow-up evaluations that go beyond just a few weeks or 3–4 months after the training, (ii) unobservable background variation in school environments may swamp potential treatment effects, (iii) training may only lead to far-transfer effects in specific subject pools such as young children. Other difficulties involve (iv) choosing an appropriate control group, (v) using or developing appropriate age-adjusted outcome measures, and (vi) sample size issues.

We tackle these challenges with a randomized controlled field experiment—described in more detail below—in a sample of 572 typically developing school children in the first grade of primary school. We focus on the training of relatively young children at age 6–7 years because evidence from

¹ We use the terms “cognitive” and “noncognitive skills” in the way they are typically used in economics (e.g. by Heckman, Stixrud and Urzua (2006)). Here, characteristics that are part of and related to IQ measures are considered as “cognitive” while character traits such as the ability to persist, to motivate oneself, self-control and self-regulation (e.g., to inhibit pre-potent impulses) are labelled as “noncognitive”. Of course, from a psychological perspective, both types of skills have cognitive components. For simplicity, we stick to the convention established in the economics literature but describe our skill measures in great detail so that the reader knows exactly what skills we are measuring.

economics indicates that training programs for youths in their late adolescence or young adulthood may be less effective than for young children (Cunha et al. 2006; Heckman 2006). Young children have higher brain plasticity, which may increase the chances of generating positive far-transfer effects (Heckman 2006; Constantinidis and Klingberg 2016; Klingberg 2016; Almond, Currie and Duque 2018). In contrast to most other working memory training studies in typically developing children, we track children’s outcomes for longer than (just) 3–5 months after the training. Specifically, we measure outcomes also after 6 and after 12–13 months and we examine whether the training has an effect on children’s school trajectory 3–4 years later.

In our study, 31 school classes were randomly assigned to a treatment group (16 classes) or a control group (15 classes). Since we randomized within schools, we are able to control for unobservable background variation in school environments via school fixed effects. The children in the treatment group participated in a daily (one lesson per school day) computer-based adaptive WM training over a period of five weeks. We find not only substantial near transfer effects on WM capacity that emerge right after the five-week training period and last throughout all evaluation waves; we also find far-transfer effects on several important skills—geometry, reading ability, a measure of fluid IQ, children’s ability to inhibit pre-potent impulses, and teacher-rated self-regulation ability.² Interestingly, for all these far-transfer abilities there is no significant treatment effect shortly after the training, i.e., the far-transfer effects do not emerge in the short term. Instead, they show an increasing pattern over the course of several evaluation waves and are typically highest in the last wave (after 12-13 months) with effect sizes between 0.24 and 0.38 standard deviations. These effects are sizeable in view of the intervention’s intensity (25 school hours) and low financial costs (about US\$ 300 per child in total).³

One important aspect of our field experiment is that the WM training was embedded into the normal school routine and was introduced like any other new lesson or sequence of exercises that children experience during a school year. Thus, the children in the treatment group did not know that they were part of an experiment. The five-week WM training took place during one of the first two morning lessons during which children typically have math or German classes. This means that the children in the treatment group missed 25 school lessons relative to the children in the control group, who participated in their normal math and German lessons. Our treatment effects therefore already incorporate the opportunity cost of the lost school lessons. This means that the children in the treatment group seem to have experienced a net benefit from the WM training because the training did not reduce any outcome measure but significantly improved the children’s skill level in several dimensions. This interpretation is further corroborated by the finding that 3–4 years after the training the treatment group had a 16 percentage points higher probability of entering the academic track (called *Gymnasium*) of secondary school. In Germany, the choice of the secondary school track after the 4th grade in primary school is one of the most decisive educational choices for a child. This decision typically has a large

² With the exception of teacher-rated self-regulation abilities all our outcome measures are based on objective, computer-based tests with auditory instructions, supervised by staff that was blind to the treatments. In addition, we validate teachers’ assessments of children’s self-regulation ability with other objective measures.

³ For more details, see section VI.

influence on the probability of earning a high school (i.e., Gymnasium) degree and thus on the later university enrollment and adult labor market outcomes.⁴

Our paper is related to the literature on the role of children’s cognitive and noncognitive skills in human capital formation. Research in this area has established that not only cognitive but also noncognitive skills have an important influence on individuals’ life outcomes in terms of education, income, and health (Cunha et al. 2006; Heckman, Stixrud and Urzua 2006; Conti and Heckman 2010; Moffitt et al. 2011; Duckworth et al. 2012; Duckworth and Carlson 2013). Furthermore, research (reviewed in Borghans et al. (2008), Cunha and Heckman (2009), and Almond and Currie (2011)) has focused on the determinants of children’s cognitive and noncognitive skills and has identified the early family environment and associated parental investments, the school environment, and early health shocks as important determinants of adolescent and adult human capital. In addition, researchers have started to design interventions to boost cognitive and noncognitive skills and have conducted randomized controlled trials to measure the interventions’ causal effects. This literature examined the general role and malleability of (i) children’s “growth mindset”, i.e., an optimistic belief about the role of effort in individuals’ success (Dweck 2006; Yeager et al. 2014; Sisk et al. 2018; Yeager et al. 2019), (ii) children’s perseverance and patience (Duckworth et al. 2007; Duckworth 2011; Alan and Ertac 2018; Alan, Boneva and Ertac 2019), and (iii) children’s trust and social preferences (Kosse et al. 2020; Cappelen et al. forthcoming).

Our paper differs from these studies by focusing on different outcome measures and by choosing an intervention that has rarely, if at all, been considered by economists as a potential mechanism for changing children’s cognitive and noncognitive skills: working memory (WM) capacity. Note that WM capacity is *not* just short-term memory, i.e., the ability to store information in the short term. WM capacity also involves the ability to process information in the presence of distracting impulses and competing information that is not conducive for the individual’s goal. This is the reason why WM capacity may also be a basis for impulse control and self-regulation.

The literature on WM training in typically developing children has mostly measured the impact of training only immediately after the training or a few weeks or months after the training. There are, however, strong reasons to believe that detecting far-transfer effects to more complex skills requires follow-up evaluations that leave more time for far-transfer effects to develop. Cunha and Heckman (Cunha and Heckman 2007; Cunha, Heckman and Schennach 2010), for example, have pioneered and provided supporting evidence for the view that higher skill levels at earlier stages positively affect skill formation at later stages due to ‘self-productivity’ (skills attained at one stage augment the skills attained at later stages) and ‘dynamic complementarity’ (skills produced at one stage raise the productivity of

⁴ Dustmann (2004) finds that individuals with a degree of the academic track of secondary school (*Gymnasiumabschluss*) earn on average 54-73% higher wages at labor market entrance than those with a lower secondary school degree (*Hauptschulabschluss*, earned after 9th grade), and 22-34% higher wages than individuals with an intermediate degree (*Realschulabschluss*, earned after 10th grade).

investment into skills at subsequent stages).⁵ This is the reason why we evaluated outcomes not only shortly after the training but also 6 and 12–13 months after the training. Our findings on the time path of treatment effects corroborate the view that far-transfer effects need time to develop: in all cases in which we eventually document a significant far-transfer effect, the effect is rising over time, but in none of these cases the effect is significant already shortly after the training. However, after 6 months we observe a far-transfer effect on geometry skills, Raven’s IQ and teacher-rated self-regulation skills, and after 12–13 months we observe, in addition, a far-transfer effect on reading and the ability to inhibit pre-potent impulses.

Our paper is also related to the literature in psychology and education science that examines whether WM training (and other forms of cognitive training) lead to far-transfer effects in children (for an early contribution see Klingberg et al. (2005); for reviews see Diamond and Lee (2011) and Diamond and Ling (2020)). A considerable share of this literature focusses on children with disorders or very low WM capacity, who are known to be disadvantaged (e.g., Klingberg et al. (2005); Roberts et al. (2016)). Our paper focusses instead on typically developing children. A recent review study by Sala and Gobet (2020) lists a number of studies that examine this question but only very few of them embed the WM training into the normal school routine and compare the effect of WM training to that of normal school lessons. In addition, the outcome measures are typically taken immediately after the training or only a few months after the training. And finally, none of them measures the impact of the training on high stakes school career choices.

We believe that our approach has the advantages that (i) the children in the control group are actively engaged in their normal school lessons, i.e., we have an active and natural control group, (ii) the children in our study are not aware of being part of an experiment because the training was introduced like other new topics during normal school teaching, (iii) we can also examine a question of high policy relevance, namely whether WM training provides additional benefits or costs for the children relative to normal school lessons, and (iv) we have short- and longer-run outcome measures that enable us to study how the treatment effect evolves over time. To our knowledge, there are only two other studies (St Clair-Thompson et al. 2010; Rode et al. 2014) that implemented WM training into the normal school routine such that the effects of training relative to normal school lessons could have been assessed. Unfortunately, these two studies experienced large attrition already after a few months, and/or did not have long-term follow-up measurements.⁶ In the light of our finding that many treatment

⁵ Several authors in the psychology and education science literature (Holmes, Gathercole and Dunning 2009; St Clair-Thompson et al. 2010; Nutley and Soderqvist 2017) have also pointed out that, while near-transfer effects of WM training to untrained WM tasks may happen in the short run, training-induced improvements in WM capacity need time to affect far-transfer skills.

⁶ In St. Clair-Thompson et al. (2010) 254 children from ten classes and five different schools participated in the experiment. In each school one class was assigned to the treatment the other to the control group but it is not mentioned in the paper whether class assignment was random. Outcome measures are taken immediately after and 5 months after the training but the attrition rate for the 5-month measures was 60–70 percent. No outcome measures beyond 5 months are taken and the study does not control for school fixed effects and spillover effects within classes (i.e., no clustering at the class level). The paper finds no far-transfer effects on reading, arithmetic and other types of mathematics. In Rode et al. (2014), 11 classes of 3rd graders from five schools participated in the experiment and each school contributed between one and three classes. In four of the schools at least one class

effects only become fully visible after many months, this may have severely limited their ability to discover far-transfer effects.⁷

The rest of the paper is organized as follows: Section II describes our study design, the data collection, and our outcome measures. In addition, we put forward conjectures about the effect of WM training on our outcome measures. In Section III we describe the estimation method. In Section IV we present and discuss our empirical results in detail. Section V summarizes the results and concludes the paper.

II. Study Design and Data Description

The field experiment was conducted in primary schools in Mainz, Germany, in 2013/2014 after receiving ethical approval in September 2012.

A. Participants

With the aid of the school authorities, we recruited 31 first grade classes from numerous schools in the city of Mainz, Germany, for participation in the study. Each school participated with at least two classes. Out of 599 children in these classes in November 2012, we received the consent from 580 parents (consent rate of 96.8%) for four waves of data collection (W1, W2, W3, W4) on their children. We were able to collect test data for 572 of these 580 children at baseline (W1) and shortly (i.e., 4-5 weeks) after the training (W2).⁸ Randomization was done between classes and within schools: 15 classes (279 children, i.e., 49%) were randomly assigned to the treatment group and 16 classes (293 children) to the control group. Randomization occurred within schools enabling us to control for school fixed effects. Summary statistics are reported in Table 1. About 49% of the children were male, mean age at the beginning of the year (i.e., on January 1, 2013) was 82 months (6.8 years, SD = 4.3 months). Attrition over the course of the four evaluation waves (from W1 to W4) was very low (only about 7%, with no difference between treatment and control group, see Online Appendix Section 1.1).

was in the treatment and at least one other class was in the control group. Because one school contributed only one class there was no control group in that school (which makes it impossible to control for school fixed effects), and the empirical analysis does not take into account within-class correlations by clustering on the classroom level. All outcome measures were taken within the first 8 weeks after the training. The paper finds no far-transfer effect on the Wechsler Individual Achievement Test for math and reading (measured one week after the training) and on a curriculum-based test of reading fluency (measured 8 weeks after the training), but the authors report a positive effect on a curriculum-based test of math (measured 8 weeks after training).

⁷ There are also a number of studies that implement randomized WM training for children *outside* the school context (see review by Sala and Gobet (2020)), i.e., the children know that they are part of a study. Most of these papers measure outcomes between a few weeks and three months after the experiment and none beyond 5–6 months after the experiment.

⁸ Six children completed the W1 tests slightly after the actual start of the WM training (two of them in the control group) because they were sick or absent at the original test date. Since the delays were rather small, we kept these children in the sample. Dropping them from the sample does not change our results.

B. Treatment and Control Condition

The treatment consisted of a daily WM training session lasting approximately 30 minutes, taking place during the first or second lesson of a school day over a period of 25 consecutive school days. The WM training was embedded into the classes' normal school routine. Accordingly, parental consent on their children's participation in the training was not required and thus all children in the treatment classes participated in the training. In each class, a single teacher covers (almost) all the topics that need to be taught according to the first-grade curriculum. Thus, the WM training was introduced to the children as a normal sequence of exercises by this teacher, similar to when the teacher introduces a new sequence of exercises for math, reading, or writing as required by the curriculum. Accordingly, the teacher was present during the lessons when the WM training took place, children remained in "their" classroom, and they conducted the training sessions at their usual desks. This minimizes Hawthorne or demand effects because it ensures that the children viewed the WM training simply as a usual topic of their curriculum, in which the sequential introduction of new learning content during the school year is part of normal school routine.

Table 1: Summary Statistics

Variable	Mean	Std. Dev.	Min.	Max.	N
Working memory training	0.488	0.5	0	1	572
Male	0.49	0.5	0	1	572
Children's age in months on Jan 1, 2013	82.129	4.324	72.222	101.578	572
Children's age on test day W1 (in months)	84.247	4.377	74.523	103.485	572
Children's age on test day W2 (in months)	87.288	4.355	77.745	106.706	572
Children's age on test day W3 (in months)	92.368	4.379	82.774	111.703	544
Children's age on test day W4 (in months)	99.582	4.381	90.467	118.836	531
Migration background	0.451	0.498	0	1	568
Language problems	0.247	0.431	0	1	572
Monthly HH-Net Income <750 Euros	0.023	0.149	0	1	441
Monthly HH-Net Income 750-1500 Euros	0.12	0.326	0	1	441
Monthly HH-Net Income 1500-2500 Euros	0.209	0.407	0	1	441
Monthly HH-Net Income 2500-5000 Euros	0.433	0.496	0	1	441
Monthly HH-Net Income >5000 Euros	0.215	0.412	0	1	441
Mother university degree	0.446	0.498	0	1	444
Mother vocational degree	0.423	0.495	0	1	444
Mother no professional degree	0.131	0.337	0	1	444
Academic track secondary school	0.692	0.462	0	1	393
Mixed-track secondary school	0.204	0.403	0	1	393
Non-academic track secondary school	0.104	0.306	0	1	393

The table provides socio-demographic information about our sample. The gender and age variables have been reported by the schools and are therefore available for all children. The variables 'migration background' and 'language problems' are taken from the teacher questionnaire in W1; for four children teachers reported not to know the migration background. The income and maternal education variables are taken from the parent questionnaire in W1. The information about secondary school track is taken from a survey administered to parents at age 10 of the children.

We used a commercially available WM training software⁹ providing training on different span tasks, using an age-specific user-interface, and adaptive levels of difficulty. Eight out of ten training tasks focus on visuo-spatial WM, while only two focus on verbal WM, i.e., a much larger variety of WM tasks and more training time was allocated to visuo-spatial WM training. Crucially, none of the training tasks was similar to our near-transfer evaluation tasks for measuring complex (span) WM. The teachers supervised children in each training session, and logins for the training software were user-specific and only valid during the intervention period. The children thus only had access to the training software during their dedicated training sessions (see Online Appendix Section 1.2 for further details).

WM training typically took place in the first or the second lesson in the morning. During this time, the control group teachers taught their students the usual content—primarily math and German lessons—covered in the first and the second lesson of the day for first graders in primary school. This means that subjects in the treatment group missed 25 school lessons. Therefore, even if WM training improves some math or German skills, this improvement could, in principle, fall short of the improvement that the children in the control group experienced because they received more direct training in these subjects. This paper therefore analyzes the question of which activity improves skills more. This allows us to address a question of particular importance for education policy, i.e., whether WM training during school hours is beneficial for the children. In other words, when we compare the treatment and the control group children on the various skill dimensions, we automatically take the foregone school lessons during WM training, i.e., the opportunity cost of the training, into account. This is important for an overall assessment of the desirability of WM training for a general school population of young children—the training is not without cost. In fact, it is even possible that—due to the loss of math and German lessons—the skill levels of the treatment group might end up being below those of the control group. Hence, if children in the treatment group perform better in math or reading tests despite missing math and German lessons, our results will be even stronger.¹⁰

Compliance with WM training was high in our sample. Only four out of 279 treated children finished less than 20 of the 25 daily training sessions. Since classes as a whole participated in the training, children missed a training session only when they did not attend school (e.g., for health reasons).¹¹

⁹ We used the WM training software Cogmed. Cogmed and Cogmed Working Memory Training are trademarks, in the U.S. and/or other countries, of Cogmed Inc. (www.cogmed.com).

¹⁰ The literature on WM training emphasizes the importance of so-called active control groups, i.e., the subjects in the control group should also be actively involved in a task. In our case, the active control group is involved in the normal teaching lessons. It is sometimes also argued that the best active control groups are those who perform *non-adaptive* WM training, i.e., the children are *not* exposed to increasingly challenging tasks when they have solved the less challenging ones. However, the disadvantage of non-adaptive training is that the children may become bored and demotivated if they face tasks that constitute no real challenge and that, therefore, lead to no improvements. For this reason, and because we were interested in the policy question whether WM training enables improvements relative to normal teaching lessons, our active control group is involved in normal teaching lessons that typically involve increasingly challenging material.

¹¹ We observe an average training index improvement of 20.76 points—a measure that is used by the training software itself (independent of our study) to quantify compliance and performance. Other studies have reported similar training improvements.

C. Data Collection

1. Computer-based Tests

Computer-based tests were completed by all children in four evaluation waves: at baseline, i.e., 3–4 weeks before the training (W1), shortly (i.e., 4-5 weeks) after the training (W2), 6 months after training (W3), and 12–13 months after training (W4) (for further details, see Online Appendix Section 1.3). Parents of both treatment and control children gave their consent to participate in the data collection (consent rate of 96.8%). The tests were highly standardized and developed specifically for the purpose of the present study. The entire sequence of tests was computer-based, including auditive (via headphones) explanations and comprehension checks. The test items for each evaluation wave were adjusted to the relevant age and school curriculum at the different waves. A pretest prior to W1 with a different (smaller) sample of similar aged children served to adapt the initial level of difficulty. The input devices for the tests were large touchscreens instead of computer mice because we wanted to avoid any bias arising from the fact that children in the treatment group had been working with computer mice during the WM training. The testing procedure was run by a professional data collection service. The staff administering the tests was blind to treatment conditions. Teachers were not present during the tests and did not know their content. The teachers also did not receive any information or feedback about the performance of their students in the evaluation tasks. When the children had finished all evaluation tasks in a given wave, we rewarded them for their participation with a selection of toys to ensure high motivation. These rewards were given to all children from the control and the treatment group to avoid any motivational differences between them.

In each evaluation wave, the children completed three (non-trained) WM tasks (near-transfer tasks) as well as several tasks to evaluate far-transfer effects in arithmetic, geometry, reading, fluid IQ, inhibition control, and attentional stamina.

WM capacity was measured with a verbal simple span task, a verbal complex span task, and a visuo-spatial complex span task (for details, see Online Appendix Section 1.4). The children had to recall sequences of single-digit numbers provided via headphone in the verbal simple span task (digit span). The verbal complex span task used a sequence of objects to be recalled, interrupted by the question of whether the object represents an animal or not. Thus, in the complex span task the children are confronted with a higher “distraction load” that makes it more difficult to recall the objects. The visuo-spatial complex span task consisted of a series of screens each of which showed a horizontal list of three symbols in which the child had to identify and remember the positions of the slightly deviant symbol; after a sequence of screens, the child had to recall the position of the deviant symbols in the correct order. Thus, both the verbal complex span task and the visuo-spatial complex span task clearly differ from the training tasks. We included a verbal simple span task (but not a visuo-spatial simple span task) in the set of our WM evaluation tasks because the WM training places considerably less weight on verbal compared to visuo-spatial WM. Near transfer effects may therefore be weaker for verbal WM. We included a verbal simple span task to allow us to still be able to capture these weaker effects. The three WM tasks mentioned above not only enable us to study near-transfer effects but they

also serve the purpose of examining the extent to which WM capacity mediates training-induced improvements in far-transfer tasks.

In each evaluation wave, the children also completed a set of far-transfer tasks (for details, see Online Appendix Section 1.4). Educational achievement was measured in three areas: arithmetic, geometry, and reading. We included geometry as an outcome measure because—like arithmetic and reading—it plays an important role in everyday life (e.g., orientation, reading maps, driving, and parking) as well as in various professions (e.g., construction/architecture, fashion/art design, geography, astronomy, physics, sports, etc.). In addition, we had three other far-transfer tasks that measure important aspects of fundamental skills like the ability to inhibit pre-potent responses, the ability to sustain attention, and fluid intelligence. Fluid intelligence was measured using Raven's Colored Progressive Matrices test (Bulheller and Häcker 2010), the ability to inhibit pre-potent responses was measured with the go/no-go task (Gawrilow and Gollwitzer 2008), and attentional stamina was measured using the bp task (Esser, Wyschkon and Ballaschk 2008).

In the go/no-go task the child faces a sequence of screens each of which shows an animal. For the large majority of the animals (“target animals”) the children need to push a red button on the touchscreen every time one of these animals appears on the screen. However, for one other (“non-target”) animal, that appears only rarely in the sequence of screens, the children must not push the red button (see Online Appendix Figure S10). Each screen is only shown for a short time window during which the children must decide whether to push the button and to implement the button press. Because the target animals occur much more frequently than the non-target animal and the time window during which a decision can be made is short, the children are put in the “go-mode”. In other words, the pre-potent impulse is to push the red button. A key challenge in this task is, therefore, to inhibit the pre-potent impulse when a non-target animal appears. Commission errors in this task are widely viewed as a behavioral measure of impulsivity and lack of self-control (Helmers, Young and Pihl 1995; Eigsti et al. 2006).

In the bp task the subjects see 45 randomly ordered letters during each trial and each letter is either a ‘b’, ‘d’, ‘g’, ‘h’, ‘p’, or ‘q’ (see Online Appendix Figure S11). The child has to highlight (i.e., touch) *only* the letters ‘b’ and ‘p’ on the touchscreen. Thus, in contrast to the go/no-go task the children are here not habituated to a particular behavioral response (“go”) that they must inhibit from time to time. Rather, the children have to continuously find (and touch) the letters b and p.

2. Teacher Ratings

In each data collection wave (W1–W4), teachers filled out a questionnaire containing items on children’s characteristics—such as their migration background or language problems—and teacher characteristics. We achieved a 100% return rate for the teacher questionnaire in all four evaluation waves. A key part of the teacher questionnaire is a series of questions capturing teachers’ assessment of each child’s self-regulatory abilities (for details, see Online Appendix Section 1.4). We use the

standardized sum of the standardized answer values on the individual questions as an additional outcome ('teacher-rated self-regulation') in the analysis below.

3. Secondary School Track Choice

In a follow-up survey in spring 2016, we asked parents to report their children's school track for secondary school in fall 2016. Secondary school starts at grade five, i.e., 3-4 years after the WM training when the children are 10-11 years old. Essentially, there are three different secondary school tracks available: (i) an academic track (*Gymnasium*), (ii) a mixed track (*Integrierte Gesamtschule*), and (iii) a non-academic track (*Realschule Plus*). In this particular federal state in Germany, 86 percent of the children in the academic track earn a degree that qualifies them for general university enrollment (*Abitur*), whereas only 25% percent of children in mixed-track schools achieve this (Rhineland-Palatine 2018). Within the non-academic track, students cannot earn a degree that qualifies them for general university enrollment. For children in the non-academic track, the probability of switching track is small (< 5% per year) (Bellenberg 2012). Moreover, since the early school track choice at this age has a decisive influence on the whole educational career path, it also exerts a substantial influence on later wages (Dustmann 2004). Thus, the choice of the secondary school track constitutes a major educational decision that strongly affects a child's future outcomes and life-time earnings.

D. Conjectures About the Treatment Effect on Outcome Measures

There is reason to believe that WM training may have a positive effect on performance in all educational achievement tasks we measured, but in varying degrees. Performing arithmetic tasks such as adding or subtracting several numbers requires children to store and recall "intermediate results" while performing the computations, thus requiring WM capacity. Likewise, geometry tasks, such as estimating how many times a smaller geometrical object fits into a larger one, and reading comprehension require WM capacity. In our context, it is however important to take into account that teaching time in primary school is very unevenly allocated between arithmetic and geometry; during the first grade, the curriculum requires that about 70% of the math lessons be spent for teaching arithmetic. Because the treatment subjects miss a considerable number of math lessons and because our WM training was focusing on visuo-spatial WM (see above), it seems more likely that we find positive training effects on geometry than on arithmetic skills.

With regard to reading performance, it is important to keep in mind that the children gradually learn the various letters of the alphabet during the first grade, allowing them to read and understand an increasing number of letters and words over time. We measured reading skills by a reading comprehension task that required children to understand and process all words in a sentence, and to assign meaning to the full sentence. This is obviously much more difficult when children still have problems reading single words. Moreover, correlational evidence suggests (Kibby, Lee and Dyer 2014; Nutley and Soderqvist 2017) that WM capacity does not predict word identification, but it seems to be

an independent predictor of reading comprehension once word reading ability has been acquired. This suggests an additional, independent reason—apart from the possibility that far-transfer effects generally may need time to emerge—for why WM training effects in our reading task may only emerge over time.

WM capacity has also been shown to be strongly correlated with fluid intelligence as measured, for example, by the Raven’s matrices task—a task that requires visuo-spatial WM but is nevertheless very different from pure WM tasks because it requires (i) reasoning in novel situations without prior knowledge, (ii) the ability to generate high-level schemata in order to handle complexity, as well as (iii) the ability to absorb, recall, and reproduce information provided in the task (Carpenter, Just and Shell 1990; Oberauer et al. 2005; Wiley et al. 2011). Therefore, WM training may improve performance in Raven’s matrices task. However, the previous empirical literature is in sharp disagreement about whether WM training improves fluid IQ measured using Raven's Matrices tasks (Au et al. 2015; Melby-Lervag, Redick and Hulme 2016).

WM capacity is not simply the ability to store and recall items. Instead, working memory is a form of “executive attention”, that is, the ability to actively maintain task-relevant and suppress/inhibit task-irrelevant information (Engle 2002). WM capacity is thus predicted to enhance the ability to avoid distraction, which is consistent with the evidence showing that individuals with low WM capacity are less able to suppress salient distractors (Gaspar et al. 2016). Based on this account, children who undergo WM training should be better able to avoid commission errors in the go/no-go task because the children in this task almost always see symbols that require them to press a button within a very short time interval, placing them in the “go-mode”. Occasionally, however, a “no-go” symbol is shown that requires them to *refrain* from pressing the button. In this view, the frequent display of “go-symbols” distracts individuals and makes it difficult for those with low WM capacity to maintain the goal and provide the appropriate behavioral response associated with the “no-go-symbols”.¹²

Finally, we measure children’s attentional stamina with the so-called bp task. It is an open question whether WM training improves performance on the bp task but we nevertheless thought that it is interesting to know whether it does. In fact, according to one theory of WM capacity, there are reasons to believe that WM training increases performance in the go/no-go task but not in the bp task. We will come back to this theory when we interpret our empirical results.

Because we conjectured that WM training may enhance inhibition control, and given that inhibition control is a key component of self-regulation ability, we also thought that we may find a positive treatment effect on children’s overall self-regulation behavior in the classroom as assessed by their teachers.

¹² Redick et al. (2011) hypothesize and show that subjects with low WM capacity display lower performance in the go/no-go task – a finding that is consistent with our conjecture. Other approaches also stipulate a close association between inhibitory abilities and working memory capacity (e.g., L. Hasher, C. Lustig, A. R. A. Conway, Inhibitory Mechanism and the Control of Attention. 2007). They view variation of inhibitory efficiency as the main driver of variation in cognitive performance and working memory capacity.

Finally, in case we find that WM training increases academic performance and some of the other important skills, it might be possible that training also affects secondary school track choice positively because that choice is presumably influenced by children’s academic skills, their fluid IQ and their self-regulation skills.

III. Empirical Results

To estimate the treatment effect of WM training, we regress outcome scores measured after the training (W2–W4) on a treatment indicator and a vector of control variables (including school fixed effects, age, gender, etc.—see Online Appendix Section 1.5 for details). All outcome scores are standardized within each evaluation wave to mean = 0 and standard deviation = 1. We control for the pre-training baseline level of the respective outcome score in our regressions. Thus, instead of identifying how WM training changes individuals’ outcome scores between pre- and posttreatment waves (i.e., using the difference-in-differences estimator), we estimate how the training changes outcome *levels* and control for the baseline level of the respective outcome. 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 (Frison and Pocock 1992; McKenzie 2012). Finally, in order to allow for interdependence of observations within school classes, standard errors are clustered at the classroom level. In our robustness analysis we also apply the Romano-Wolf stepdown procedure to control for multiple hypothesis testing (Romano and Wolf 2005; Romano and Wolf 2016)—a technique that is increasingly used for large-scale intervention studies (see, for example, Cunha, Heckman and Schennach (2010), Campbell et al. (2014), Gertler et al. (2014))—and, simultaneously, we control for potential biases that may arise when the number of clusters is relatively small with the BRL (biased-reduced linearization) correction method (Bell and McCaffrey 2002).

A. Sample Balance

To examine whether randomization led to a balanced sample across treatment and control group in terms of socio-economic characteristics, we regress various socio-demographic characteristics (gender, age, migration background, as well as parental income and education) measured prior to the treatment (W1) on the treatment indicator (see Table 2). In these regressions, a significant coefficient related to the treatment dummy would indicate that the sample is not completely balanced between the treatment and the control group with regard to the socio-economic characteristics. The results show that the treatment coefficient in all regressions is close to zero and insignificant, indicating that there were no significant imbalances between treatment and control group with respect to these variables. The estimations in Tables 2 are based on a linear probability model with standard errors clustered at the classroom level.¹³

¹³ If we use probit models instead, the results are basically the same. When looking at pairwise correlations instead of regressions, findings are very similar as well.

As a further sample balance check, we regressed standardized outcome test scores at baseline (i.e., test scores measured prior to the treatment in W1) on the treatment dummy, school fixed effects, and the same control variables that are included in the main estimations of the treatment effect. Table 3 below shows that with the exception of the baseline score for the verbal complex span task, none of the coefficients related to the treatment dummy is significantly different from zero, indicating that for all other baseline test scores there is no evidence for significant imbalances between treatment and control group. With regard to the possible imbalance in the baseline score of verbal complex span, we have to take into account that we conducted a total of 15 imbalance test regressions. For this reason, we further examined the issue by adjusting p-values for multiple hypothesis testing and applying the biased-reduced-linearization clustering method (which accounts for small numbers of clusters). This then yields a p-value of 0.332 for the verbal complex span outcome, suggesting no significant difference between the treatment and control group once we account for the number of tests conducted. In addition, we would like to mention that we control for the baseline tests scores in W1 in all our regressions that measure the treatment effect of WM training on outcome scores in W2–W4.

Table 2: Sample Balance: Regressing Socio-demographic Characteristics on the Treatment Indicator

	Male (1)	Age (2)	Migration Background (3)	Language Problems (4)	Income Eur 2500+ (5)	Mother Univ Degr (6)
Working memory training	-0.013 (0.034)	-0.772 (0.501)	-0.128 (0.081)	-0.035 (0.071)	0.084 (0.087)	-0.051 (0.072)
N	572	572	568	572	441	444
R squared	0.028	0.040	0.129	0.127	0.213	0.134

The results are based on least squares models including school fixed effects. Standard errors in parentheses are clustered at the classroom level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The coefficients for ‘working memory training’ in the first row and the associated standard errors indicate whether there are significant imbalances between the treatment and control group with respect to the socio-demographic characteristics described in the column titles. In every column the coefficient for working memory training is small and insignificant. The sample 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. The samples in columns 5 and 6 are smaller because the dependent variables are taken from the parent questionnaire, which has not been answered (completely) by all parents.

Table 3: Sample Balance: Regressing W1 Baseline Test Scores on the Treatment Indicator

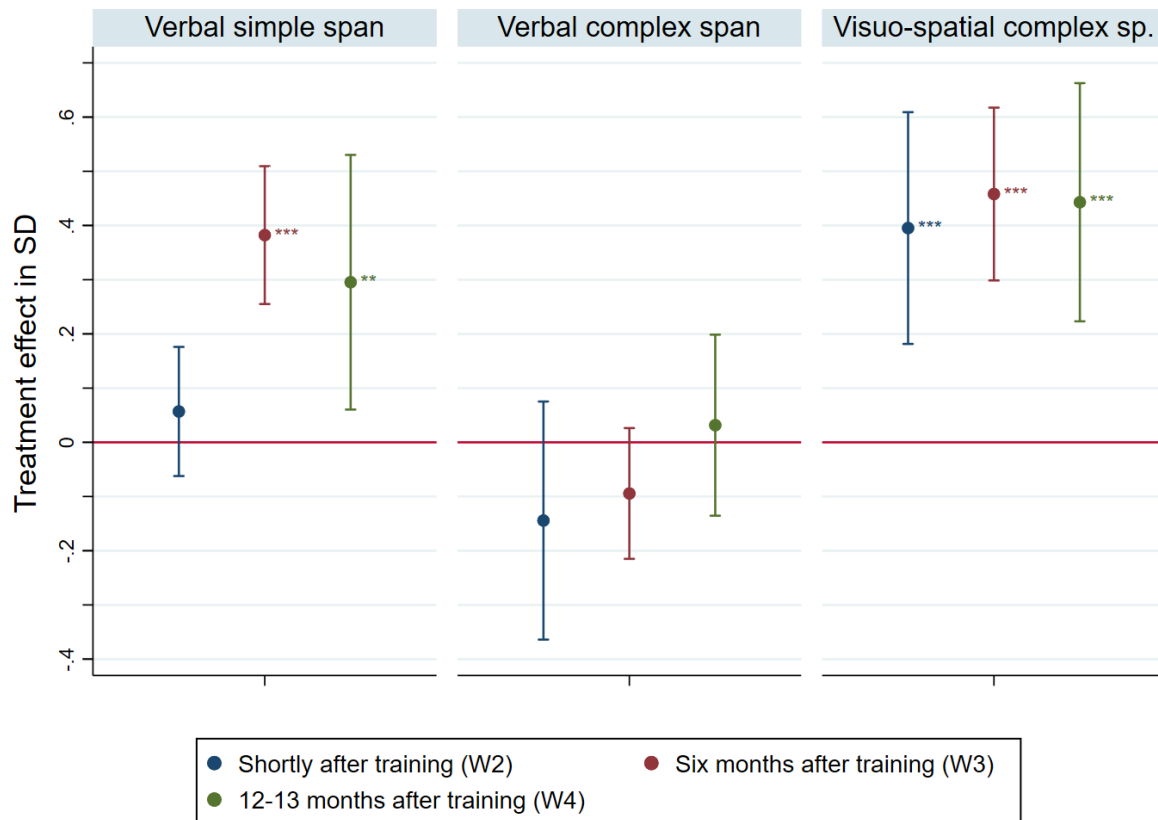
	Verbal simple span (1)	Verbal complex span (2)	Visuo-spatial complex span (3)	Geometry (4)	Arithmetic (5)	Reading (6)	Raven's IQ (7)	Go/No-Go task (8)	bp task (9)
Working memory training	-0.079 (0.104)	0.274** (0.128)	-0.160 (0.101)	0.082 (0.121)	0.102 (0.102)	0.136 (0.182)	0.060 (0.102)	-0.104 (0.136)	0.133 (0.128)
N	569	566	567	568	549	567	568	567	565

The results are based on least squares models including school fixed effects and further controls (see Online Appendix Section 1.5 for details). All outcome scores are standardized to mean = 0 and SD = 1. Standard errors in parentheses are clustered at the classroom level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. 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 'working memory training' (except the one for verbal complex span) are close to zero and insignificant at the 5% level, i.e., there is no evidence for significant imbalances between treatment and control group for these outcome measures. Because the testing for imbalances involved many hypothesis tests, we further check whether the significant coefficient for verbal complex span survives multiple hypothesis correction. If we adjust the p-value for multiple hypothesis testing, the coefficient for verbal complex span turns insignificant ($p = 0.332$). Note that we control for the baseline (i.e., W1) scores of all outcome variables when we estimate the treatment effect of working memory training.

B. Treatment Effect on Computer-based Test Outcomes

To estimate the effect of WM training, we regress outcome scores measured shortly after the training (W2), 6 months after the training (W3), and 12–13 months after the training (W4) on the treatment indicator. The estimated treatment effects on near-transfer measures (WM capacity) are presented in Figure 1 and Supplementary Table S1 in the Online Appendix. We find significantly positive treatment effects for the visuo-spatial complex span task in all three post-treatment waves with an effect size (d) of 0.40–0.46 SD ($p = 0.00004$ – 0.006). We also find a significantly positive training effect on performance in the verbal simple span task of $d = 0.38$ SD ($p = 0.000008$) in W3 and $d = 0.30$ ($p = 0.015$) in W4. We do not find any significant treatment effect for performance in the verbal complex span task. The stronger effect of training on visuo-spatial WM compared to verbal WM is plausible, as the training focused primarily on visuo-spatial WM.

Figure 1: Treatment Effect on Working Memory Capacity

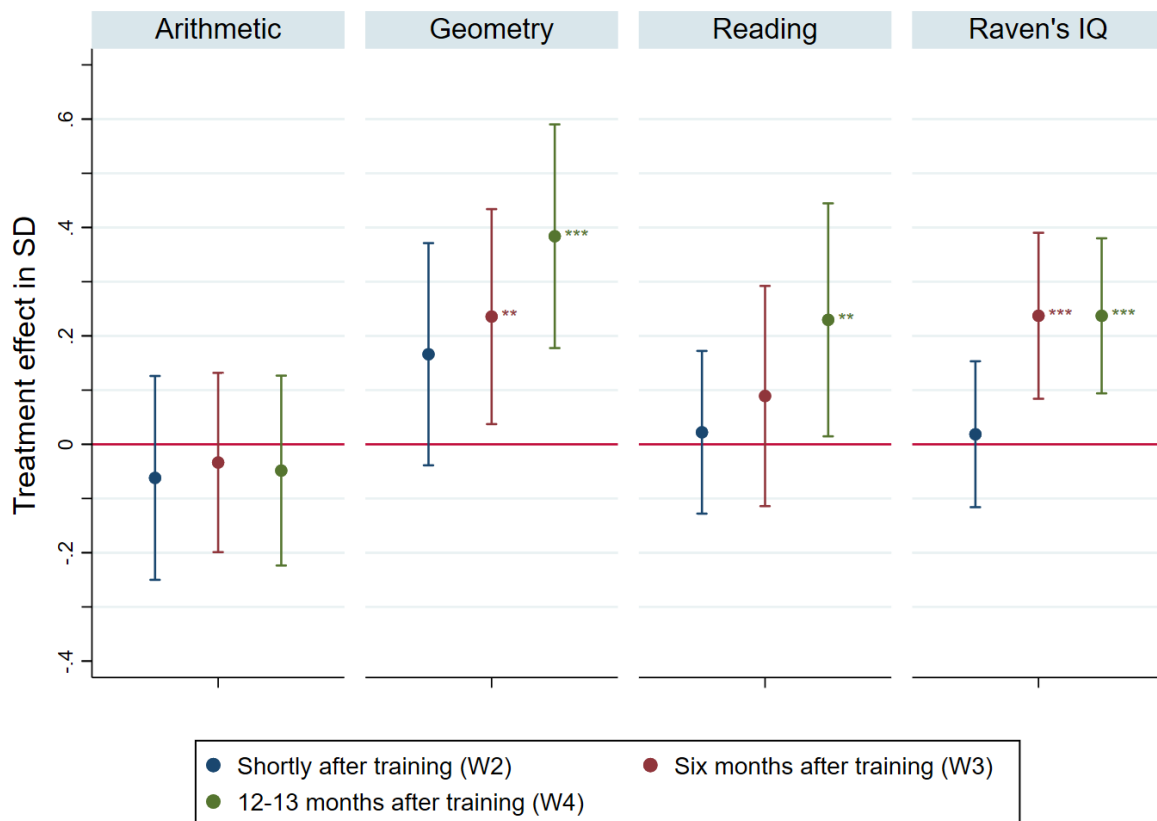


The dots show the point estimates (as fractions of a standard deviation) of how WM training changes the performance in the three working memory tasks (indicated in the subfigure title) relative to the control group. The bars indicate the 95% confidence intervals. All estimates are based on least squares models controlling for school fixed effects, pre-treatment outcome scores, and further controls (see Online Appendix Section 1.5 for details). The confidence intervals and the associated significance statements are computed based on the clustering of standard errors at the classroom level. Stars refer to significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

The far-transfer effects of WM training on educational outcomes—arithmetic, geometry, and reading—and Raven’s IQ measure are reported in Figure 2 and Supplementary Table S2 in the Online Appendix. While there is no treatment effect on arithmetic in all three post-training waves, we find an effect on geometry skills that is increasing over time. The effect size $d = 0.17$ in W2 is not yet significantly different from zero ($p = 0.108$),

but the effect size increases in W3 and W4 to $d = 0.24$ and $d = 0.38$, respectively, with significance levels of $p = 0.021$ in W3 and $p = 0.001$ in W4. Thus, it seems that WM training had a positive and increasing long run effect relative to the normal school curriculum on geometry skills but not on arithmetic skills. This difference in treatment effects may reflect the fact that children in the control group received considerably more arithmetic teaching during the intervention. In addition, the difference is also consistent with the fact that training focused on and improved visuo-spatial WM capacity more strongly than verbal WM capacity.

Figure 2: Treatment Effect on Arithmetic, Geometry, Reading and Raven’s IQ



The dots show the point estimates (as fractions of a standard deviation) of how WM training changes performance in arithmetic, geometry, reading, and Raven’s IQ task, respectively, relative to the control group. The bars indicate the 95% confidence intervals. All estimates are based on least squares models controlling for school fixed effects, pre-treatment outcome scores, and further controls (see Online Appendix Section 1.5 for details). The confidence intervals and the associated significance statements are based on the clustering of standard errors at the classroom level. Stars refer to significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

The effect sizes for reading are generally lower than for geometry but they are also rising over time and become significant in W4. There is no positive effect on reading shortly after the training, but we observe a larger, yet still insignificant effect in W3 and an effect size of $d = 0.23$ at $p = 0.037$ in W4. This rising pattern for the reading outcome is consistent with the view (Nutley and Soderqvist 2017) that WM capacity plays a smaller role for reading comprehension when children are still struggling to understand words, but eventually becomes relevant for reading comprehension when word identification has progressed sufficiently.

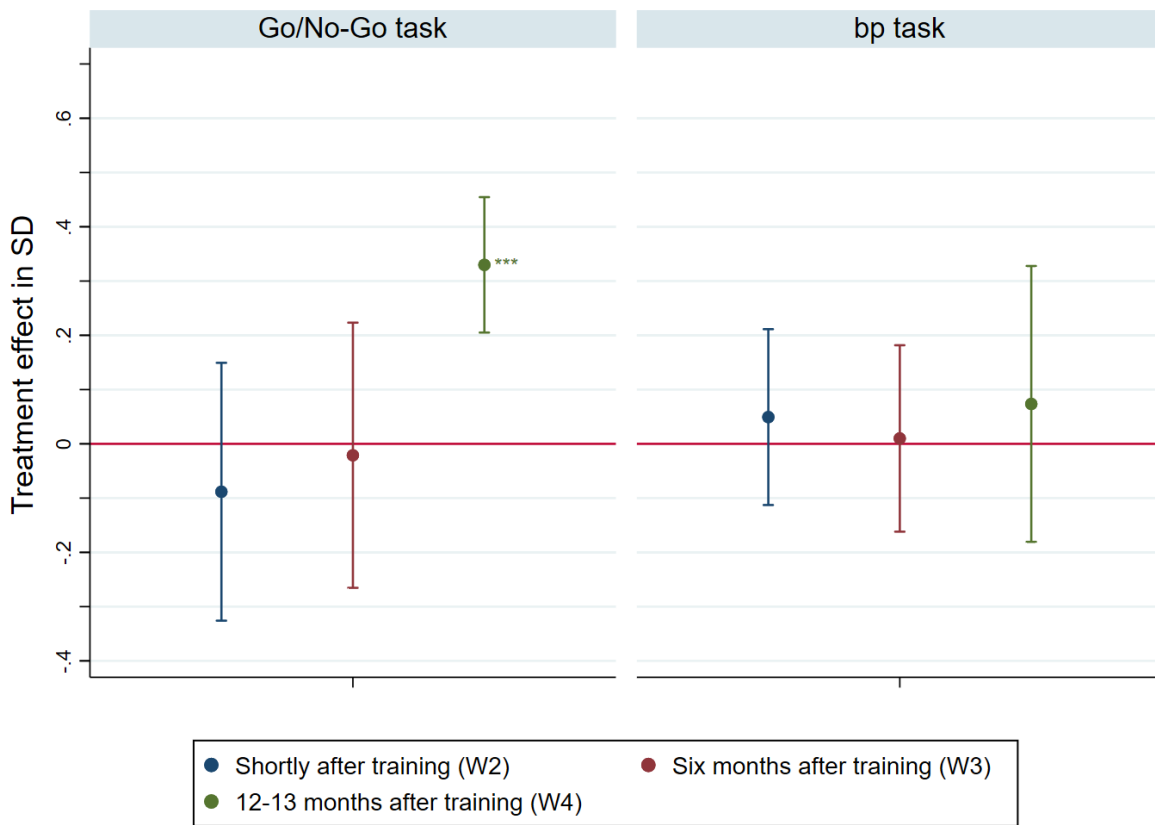
We find a significant treatment effect for Raven’s Colored Matrices task six months ($d = 0.24$, $p = 0.004$) and 12–13 months after the training ($d = 0.24$, $p = 0.002$). We emphasize that this finding does not mean that WM training increased all dimensions of fluid intelligence, as some research indicates that “only” 64% of the variance in performance in a Raven’s task is attributable to general fluid intelligence (Jensen 1998). However, the Raven task measures important dimensions of fluid intelligence which require WM (Carpenter, Just and Shell 1990) and its deployment in novel situations (Wiley et al. 2011).

It is also important to mention that *none* of the treatment effects in geometry, reading, or Raven’s IQ measure are driven by a *decline* in the performance of the control group. Due to cognitive maturation over the course of one year, both the treatment and the control group increased their performance over time. Therefore, the treatment effects are due to a differentially larger increase in performance in the treatment group.

Finally, we turn to the effects of WM training in the go/no-go task and the bp task (Figure 3 and Supplementary Table S3 in the Online Appendix). We find that WM training improves children’s inhibitory abilities measured in the go/no-go task. We measure inhibitory performance formally by multiplying children’s standardized number of commission errors with -1, i.e., a reduction in commission errors shows up as a numerical increase in this performance measure. Figure 3 (and Supplementary Table S3 in the Online Appendix) indicate a highly significant reduction in commission errors in the treatment relative to the control group in W4 ($d = 0.33$, $p < 0.0001$). If, instead of commission errors, we use the standardized (i.e., z-scored) d' -measure of performance in this task—which subtracts the standardized fraction of commission errors in the no-go trials from the standardized fraction of correct responses in the go trials—we also find a significant performance effect in W4 ($d = 0.48$, $p < 0.0001$), see Supplementary Table S3. Interestingly, while we observe no treatment effect on commission errors (and the d' -measure) in W2 and W3, we observe a weakly significant treatment effect on performance in terms of a reduction in response times in W2 ($d = 0.23$, $p = 0.053$) and W3 ($d = 0.37$, $p = 0.094$). Thus, although the children in the treatment group did not make fewer mistakes in W2 and W3 they were quicker in delivering their responses (without increasing their mistakes) in these evaluation waves.

Overall, these data patterns suggest that, similar to the case of geometry, reading, and Raven’s IQ measure, far-transfer effects slowly emerge over time in the go/no-go task. Note also, that the treatment effects in the go/no-go task are due to a differentially larger increase in the performance of the treatment group relative to the control group in terms of shorter response times or fewer errors. In contrast to the results in the go/no-go task, we cannot detect a training-related improvement in performance in the bp task. In fact, the time profile of the treatment effects is completely flat and close to zero suggesting that WM training does not affect attentional stamina. In our interpretation section below, we will discuss the differential effect of WM training in the go/no-go task and the bp task because this difference may inform theories of WM capacity.

Figure 3: Treatment Effects in the Go/No-Go Task and the bp Task



The dots show the point estimates (as fractions of a standard deviation) of how WM training changes the performance in the go/no-go task and the bp task relative to the control group. The bars indicate the 95% confidence intervals. All estimates are based on least squares models controlling for school fixed effects, pre-treatment outcome scores, and further controls (see Online Appendix Section 1.5 for details). The confidence intervals and the associated significance statements are computed based on the clustering of standard errors at the classroom level. Stars refer to significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

C. Treatment Effect on Choice of Secondary School Track

Our finding that WM training has a positive treatment effect on several far-transfer outcomes relevant for the school context suggests the possibility that it might affect children's further school career. As mentioned previously, one of the most consequential school track choices in the German education system is whether the children enter the academic track (called *Gymnasium*) of secondary school. This choice is typically taken around age 10, i.e., 3 years after the children received the WM training.

Controlling for the same set of variables as for the other treatment effects, we find indeed that children in the treatment group are roughly 16 percentage points more likely to be enter the academic track of secondary school relative to children in the control group (Table 4, column 1). If we estimate the treatment effect with a probit model instead of a linear probability model (Table 4, column 2), the result is very similar—the children in the treatment group are again roughly 15 percentage points more likely to be enrolled in the academic track of secondary school. If we take the full range of secondary school choices (academic track, mixed track, non-

academic track¹⁴) into account, we again find a sizeable positive treatment effect on enrollment in the academic track (columns 3 and 4). Column 4 of Table 4 also indicates that the increase in academic track enrollment by roughly 14 percentage points is due to a decrease in mixed track enrollment by roughly 7 percentage points and a similar decrease in non-academic track enrollment.

Table 4: Treatment Effect on Secondary School Choice at Age 10

Treatment Effect of WM training on choice of	(1) OLS	(2) Probit	(3) OLS cat var	(4) Ordered Probit	(5) Inverse Prob Weighting
Academic track school	0.157*** (0.050)	0.148*** (0.045)	0.221*** (0.078)	0.136*** (0.046)	0.165*** (0.052)
Mixed-track school				-0.067*** (0.025)	
Non-academic track school				-0.069*** (0.023)	
N	393	393	393	393	378

Column 1 reports the effect of the treatment on the probability of being enrolled in an academic track secondary school based on a least squares model. Column 2 reports the marginal effect of the probit estimate of the treatment effect on the same dependent variable as in column 1. Column 3 reports the least squares effect on a categorical dependent variable. This variable takes on value 1 if the child is enrolled in a non-academic track school (*Realschule Plus*), value 2 if the child is enrolled in a mixed-track secondary school (*Integrierte Gesamtschule*), and value 3 if the child is enrolled in an academic track school (*Gymnasium*). Column 4 reports the marginal effects of the ordered probit estimates of the treatment effect on the same dependent variable as in column 3. Column 5 reports a similar estimation as in column 1 but accounts for attrition by applying inverse probability weighting. The weights are calculated for groups defined based on migration background, high/low academic performance (math and reading performance), and high/low cognitive performance (WM capacity and Raven’s IQ). All models include school fixed effects and further controls (see Online Appendix Section 1.5 for details). Standard errors in parentheses are clustered at the classroom level. * p<0.10, ** p<0.05, *** p<0.01.

As we measure the secondary school track enrollment more than three years after the WM training, we naturally observe some attrition. This is due to reasons such as families moving away from the city of our study or when the parents do not answer the long-run follow-up questionnaire. Importantly, however, we do not observe a systematic difference in attrition between treatment and control group. In the treatment group, we still can collect data of 73.34% of the sample in W4 and in the control group we have data of 74.63% of the sample in W4. The absence of systematic attrition differences between treatment and control group is also suggested by the following regression analysis: If we regress participation in the long-run follow up questionnaire for the school track choice on a treatment dummy, school fixed effects and further controls we find that the coefficient related to the treatment dummy is close to zero and insignificant (p = 0.337).

An alternative way to check for systematic attrition is to estimate inverse probability weighting models. We computed the weights based on (i) the migration background of the children, (ii) their academic performance in W1 (reading, arithmetic, geometry), and (iii) their cognitive performance in W1 (the three

¹⁴ In Germany, the non-academic track is called “*Realschule Plus*”, the mixed track is called “*Integrierte Gesamtschule*”, and the academic track is called “*Gymnasium*”.

working memory tests and IQ-score). The result of this model (shown in column 5) also indicates that the WM training increases academic track enrollment by roughly 17 percentage points.

To gauge the size of our effect on school track choice, consider the relationship between parental education and school track choice for the control group: for children whose mother has a university degree, 86 percent chose the academic track; for those whose mother does not have a university degree, the number is 54 percent. Thus, the 14–17 percentage point increase in academic track enrollment is of substantial size when compared with this socioeconomic gap.

Why do we find such a large treatment effect of WM training on academic track enrollment? A possible reason might be that due to self-productivity in the process of skill formation the increase in WM capacity, educational skills, and inhibitory skills during the first year may have triggered a self-reinforcing cycle. Another reason could be that we have treated complete classes (class-wise randomization) which might have led to positive peer group effects within the treated classes.

D. Heterogenous Treatment Effects?

Do disadvantaged children benefit particularly strongly from WM training? Existing work has raised this question and remains inconclusive (Katz and Shah 2016; Roberts et al. 2016). We examined the heterogeneity of treatment effects with regard to initial WM capacity by including a dummy variable for the children who are below the 25th percentile in the distribution of WM capacity at baseline (W1), and by interacting this dummy variable with the treatment dummy (see Online Appendix Section 1.9 and Tables S4–S6). The results show that children with low baseline WM capacity perform substantially worse in all far-transfer outcome measures (and all data collection waves) with the exception of the bp task. However, the interaction between low WM capacity and the treatment dummy is almost never significant (with the exception of geometry in W2, where we observe a positive interaction, and the bp task in W2, where the interaction is negative). This suggests that the training effect is not systematically different for children with low WM capacity. Importantly, however, the training effect is robust to the inclusion of the low WM-capacity dummy and its interaction with the treatment dummy for all outcome variables for which we previously found a significant treatment effect.

IV. Robustness Checks

We have several outcome measures and we examine the impact of WM training on them in three post-treatment evaluation waves. For this reason, we deal with the issue of multiple hypothesis testing below. 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 WM training over time on all those variables for which we ultimately find a significant treatment effect; and furthermore, we observe insignificant treatment effects (and small point estimates) across all evaluation waves in those cases in which the treatment had no impact (i.e., arithmetic and bp task). If the observed

significant effects were simply due to randomness and did not reflect true treatment effects, we would expect a more irregular pattern.

Nevertheless, it makes sense to check the robustness of our findings with respect to multiple hypothesis testing (Romano and Wolf 2005; Romano and Wolf 2016) and we also combine this with the BRL (biased-reduced linearization) correction method that accounts for potential biases in the estimation of standard errors when the number of clusters is relatively small. When we apply these two robustness checks, we still find significant treatment effects on near-transfer outcomes (WM capacity) and on geometry, Raven's fluid IQ measure, and the go/no-go task, while the effect on reading in W4 is no longer significant (see Online Appendix Table S7). Thus, the thrust of our training effects survives these checks, which lends credibility to our results.

Next, we discuss the concern that treated children might have improved their outcome scores solely because of a Hawthorne or demand type effect (Melby-Lervag and Hulme 2013). In our view, several reasons speak against this possibility. First, the WM training was embedded into the normal school routine and was introduced like any other new sequence of exercises that children experience during a school year. Thus, the children in the treatment group did not know that they were part of an experiment. In addition, both the children in the control and the treatment group participated in the test tasks, implying that participation in these tasks also cannot explain differential performance across groups. In fact, both the children in the control group and the treatment group were highly motivated in performing the tasks and reported to enjoy taking part in them (see Online Appendix Figures S13–S14). We find neither a treatment effect on the subjective effort provided in the evaluation tasks nor on the extent to which children enjoyed these tasks (Online Appendix Table S14). Second, the time pattern of far-transfer effects speaks against Hawthorne type effects because if participation in an experiment affects general motivation and expectations, then the effects should be most visible shortly after the training when motivation and expectation effects are still fresh. In fact, however, we observe no significant far-transfer effects shortly after the training; instead the effects only arise after 6 or 12–13 months. Finally, the specificity and plausibility of the pattern of our results across tasks speaks against Hawthorn type effects. Hawthorn type effects should rather lead to a general and not a specific change in performance. For example, general Hawthorn effects should induce effects across all outcome measures but we observe no treatment effects in verbal complex span, arithmetic and the bp task. Thus, taken together, Hawthorne type effects are unlikely to be the source of the observed treatment effect patterns.

We also conducted a robustness check related to the use of computers in school. During the computer-based WM training period, the children in the treatment group naturally used computers more frequently than the children in the control group. Based on the arguments in the previous paragraph, it is highly unlikely that this generated a Hawthorne type effect, but perhaps the teachers in the treatment group subsequently used computers more often in class and this could have had effects on the children. To examine this possibility, we asked the teachers in W3 and W4 how frequently computers were used in the classroom, and we use these data to re-estimate the relevant W3 and W4 treatment effects controlling for computer use (see Online Appendix Tables S8–S9). We find that computer usage neither significantly affects the outcome measures, nor does it change the previously estimated training effects.

Another potential concern is that some outcome scores—geometry, reading, and performance in the go/no-go task—appear censored at the upper end of the distribution (see Online Appendix Figure 12). As a robustness check, we therefore estimate the treatment effect for these outcomes using the Tobit estimator that takes this censoring into account (Online Appendix Table S10). The results, however, are robust when using this alternative estimation technique.

Finally, we perform a robustness check with respect to attrition: we re-estimate the main results reducing the sample to only those children who are still in the sample in W4 (Online Appendix Tables S11–S13), but none of the results changes when doing so. This is also consistent with the fact that attrition is generally very low across evaluation waves and not systematically different between treatment and control group.

Taken together, the evidence shows a consistent time pattern of far-transfer effects suggesting that the observed treatment effects do not simply reflect chance findings. Moreover, most of our far-transfer effects are robust to multiple hypothesis testing and the evidence also strongly speaks against Hawthorne type effects or an impact of computer use on treatment effects. Finally, neither attrition nor censored outcome measures pose a challenge for the reported results. Instead, it seems plausible that the training-induced increase in WM capacity may be a driver of the observed treatment effect—a question to which we turn in the next section.

V. Interpretation and Mechanisms

A. Working Memory Capacity, Impulse Control, and Self-regulation

Working memory capacity is more than just the ability to temporarily store information—it also involves the capacity to process information in the presence of distracting impulses that are not conducive for the individual’s goal. This is the reason why WM capacity may also be a basis for impulse control, i.e., an ability that has been termed a “noncognitive skill” in the economic literature. Our finding that WM training causes significant increases in children’s ability to inhibit pre-potent impulses in the go/no-go task supports this view.

Because impulse control and the ability to avoid goal-incongruent distractions is a crucial component of an individual’s self-regulation we also asked the teachers to assess the children’s self-regulatory abilities more broadly in a questionnaire using seven self-regulation questions like “The child has problems waiting for his/her turn” or “The child disturbs class instruction often”. A factor analysis on these items (shown in Online Appendix Table S15) reveals that teachers’ responses can be captured by one factor—the children’s overall self-regulatory ability. This measure is based on teachers’ day-to-day experience with the children and thus has an empirical base but it is also based on teachers’ subjective perception of their experiences with the children. For this reason, we validated the teacher ratings with the objective test results observed in the go/no-go task.¹⁵

¹⁵ To assess the credibility of teachers’ ratings, we computed the correlation between the performance objectively measured in the go/no-go task (averaged for each child over W1–W4)—that measures children’s inhibitory abilities, which may well be considered as one important aspect of self-regulation—and the teachers’ broader assessments of children’s self-regulation skills (again averaged over W1–W4). This is based on the idea that if teachers’ assessments contain an objective rationale, i.e., if they have a meaningful objective basis and are not purely subjective impressions,

Based on this validation of teachers' ratings, it makes sense to examine whether the children in the treatment group are rated higher in terms of broader self-regulation. We find indeed that teachers rate the children in the treatment group as significantly better on "overall self-regulation" in W3 ($d = 0.37$, $p = 0.040$) and in W4 ($d = 0.27$, $p = 0.026$). The results of the corresponding regressions can be found in the Online Appendix in Supplementary Table S16. Thus, it appears that the WM training also led to a broader improvement in self-regulatory skills. However—given the subjective nature of the teachers' assessment—we should nevertheless interpret this result cautiously.

In our view, the contrasting effects of WM training in the go/no-to task and the bp task are also interesting because they may inform theories of WM capacity. As mentioned previously, a prominent theory of working memory (Engle 2002) views WM capacity as a form of "executive attention", i.e., as the ability to maintain goal-relevant information and to suppress or inhibit goal-irrelevant information. This emphasis on the ability to maintain goal-relevant information in the presence of distracting stimuli has interesting implications. It implies, in particular, that in tasks in which it is easy to maintain a goal, WM capacity is less important compared to a task in which the maintenance of a goal is more difficult.

Recall that in the go/no-go task the children are most of the time in the "go-mode", i.e., they see most of the time a "go-symbol" after which they quickly have to press a button; but occasionally they see a "no-go symbol" and then they have to suppress the impulse to push the button. In other words, they must keep the goal of suppressing the button in their working memory despite the distracting activity of pressing the button most of the time. Thus, this task makes goal maintenance difficult implying that children with a better working memory should learn to perform better. Consistent with this idea, there is evidence showing that subjects with a better WM capacity perform better in this task (Redick et al. 2011).

In this regard, the situation is very different in the bp task, which is a letter discrimination task that reinforces the task goal in every single trial and every single sub-task within a trial. Recall that during a trial, subjects see 45 randomly ordered letters; each letter is either a 'b', 'd', 'g', 'h', 'p', or 'q' (see Online Appendix Figure S11). The child's task is to touch all letters (and *only* those letters) on the touchscreen that are a 'b' and 'p'. This letter discrimination task requires the maintenance of this task goal *for every single letter on the screen* because the subjects have to decide for each letter whether it is a 'b' or a 'p'. Moreover, at the top of the computer screen we displayed the letters 'b' and 'p' saliently so that the subjects were reminded during every trial that they had to identify and touch these two letters in the string of 45 letters. In addition, the results of our heterogeneity analyses also suggest a limited role for WM capacity in the bp task. While the children in the lowest quartile of WM capacity at baseline perform significantly worse in geometry, reading, Raven's IQ task and the no/go-task in all post-intervention evaluation waves, low WM capacity has never a significant (i.e., in W2 – W4) influence on performance in the bp task.

then we should observe a significantly positive correlation: we indeed observe significant correlations of 0.38–0.46 for the control group and 0.35–0.40 for the treatment group both for commission errors as well as for the d' measure of the go/no-go task.

Therefore, the differential impact of WM training in the go/no-go task and the bp task is consistent with theories of WM capacity that emphasize the role of WM capacity for goal maintenance in the presence of distracting stimuli.

B. Working Memory Capacity as a Mechanism

In our view the documented treatment effects on WM capacity and far-transfer outcomes—as well as the absence of treatment effects for some outcome variables (arithmetic and bp task)—make sense and can be plausibly interpreted. For example, it is very plausible that WM training has an immediate effect on visuo-spatial WM capacity (i.e., that aspect of working memory that received the most emphasis during the training), while far-transfer effects need more time to evolve—which is exactly what we observe in our data. Likewise, the finding that WM training does not increase arithmetic but geometry skills may be due to the fact that the children in the treatment group lose a much larger number of arithmetic lessons than geometry lessons and that the training emphasized visuo-spatial WM, which may well play a larger role in geometry compared to arithmetic. Similarly, visuo-spatial WM capacity is likely to be a basic prerequisite to deploy the problem-solving skill that is required to solve Raven’s IQ task.

These examples suggest that the training induced increases in WM capacity are the mechanism through which WM training caused the far-transfer effects on geometry, reading, Raven’s IQ, and the go/no-go task. To study this question in more depth and to provide a quantitative assessment of the extent to which WM capacity is the mediating mechanism, we performed a mediation analysis by applying the method described in Heckman, Pinto and Savelyev (2013), and similar applications such as Kosse et al. (2020) and Carlana, La Ferrara and Pinotti (2018). The formal details of this method are described in Online Appendix 1.5. Intuitively, the method provides us with the share of the total treatment effect of the training on each far-transfer outcome that can be explained by the training induced changes in WM capacity.

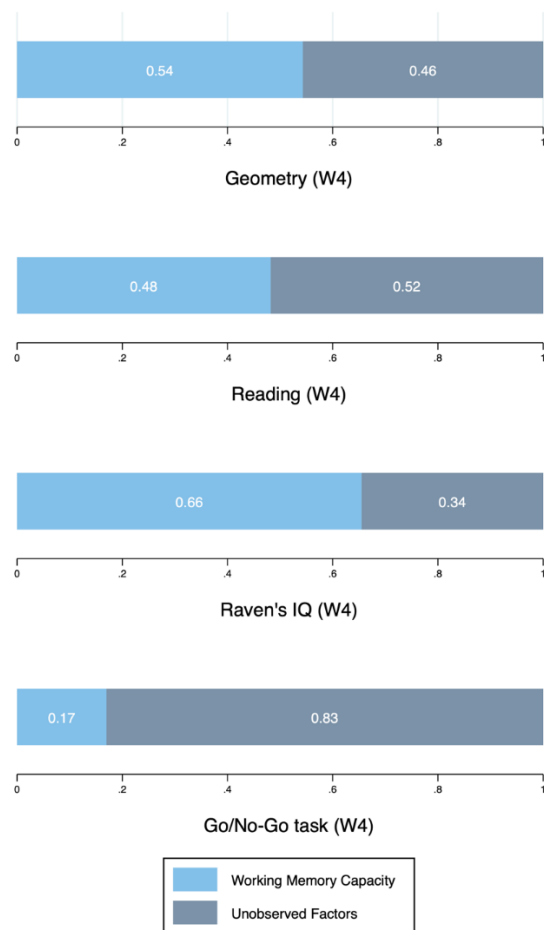
More formally, the method is based on regressions that regress each far-transfer outcome for which we found significant treatment effects (geometry, reading, Raven’s IQ, and performance in the go/no-go task) on the measured dimensions of WM capacity (simple verbal span, visuo-spatial complex span)¹⁶ and a vector of pre-program control variables X_i ; we denote the vector of coefficients that indicate the influence of the different dimensions of WM capacity on a given far-transfer outcome k , conditional on X_i , as α_{WMC}^k . The results of these regressions are shown in the Online Appendix in Table S17. In addition, the method is based on the regressions of WM capacity (simple verbal span and visuospatial complex span) on the treatment dummy and the pre-program control variables X_i (see Table S1 in the Online Appendix). These regressions provide the estimated coefficient vector β_{WMT} that measures the treatment effect of the training on the measured dimensions of WM capacity. Finally, we denote the total treatment effect of the WM training on far transfer outcome k as δ_{WMT}^k —documented in Figures 2–4 and the regression Tables S2–S3. Then, the extent to which the total treatment effect on a far-transfer outcome k is mediated by the training-induced changes in WM

¹⁶ In our analysis, we constrain ourselves to simple verbal span and visuo-spatial complex span because the WM training changed these two dimensions of WM capacity significantly.

capacity is given by $(\alpha_{WMC}^k * \beta_{WMT}) / \delta_{WMT}^k$, whereas the remaining part of the total treatment effect is given by $[1 - (\alpha_{WMC}^k * \beta_{WMT}) / \delta_{WMT}^k]$.

The results of our mediation analysis are presented in Figure 4 below. The figure shows that for geometry, reading, and Raven’s IQ measure a large part of the total treatment effect—between roughly 50% and 66%—is mediated by WM capacity. Interestingly, the mediation effect of WM capacity is much lower for performance in the go/no-go task. Perhaps this lower mediation effect of WM capacity is one reason why the training effect on performance in the go/no-go task took more time to develop. Overall, however, the most important message from this analysis is that training-induced changes in WM capacity appear to explain substantial parts of the treatment effect of the WM training on far-transfer outcomes.

Figure 4: The Relative Importance of Working Memory Capacity for the Treatment Effects on Far-Transfer Outcomes



Notes: This figure displays the estimated decomposition of the total treatment effect on those far-transfer outcome outcomes that are significantly improved by the WM training in W4 (12–13 months after treatment). For each outcome, we estimate the effect of the treatment that is mediated by WM capacity (see Online Appendix 1.5 for details). The light blue bars show the percentage of the treatment effect that is mediated by training-induced increases in WM capacity.

VI. Summary

Based on a randomized controlled trial with 572 first graders in primary schools, we found that a five-week, one lesson per school day, adaptive WM training during class improves not only children's WM capacity but also has far-transfer effects on their geometry and reading skills, their fluid IQ, and their ability to inhibit pre-potent impulses. We observe an increasing pattern of treatment effects on these far-transfer outcomes over the three evaluation waves with effect sizes ranging between 0.24 and 0.38 SD. In addition, the general pattern of our results and our mediation analysis suggest that training-induced improvements in WM capacity mediate a substantial part of the far-transfer effects. The training-induced increase in impulse control is also associated with a training-induced improvement in children's overall self-regulation ability as assessed by their teachers. When assessing the reported effect sizes for the far-transfer effects, it is useful to contrast them with effect sizes observed in other prominent—and considerably more expensive¹⁷—educational interventions such as sizeable reductions of class size (e.g., from 24 to 16 students). These interventions are typically estimated to lead to improvements in math and reading of about 0.15–0.25 SD (Angrist and Lavy 1999; Krueger 1999).

Finally, we document that the WM training has a sizeable impact on one of the most consequential school career decisions in the German school system: whether to enroll the child in the academic track of secondary school (*Gymnasium*). We find that – 3-4 years after the training – the treated children are roughly 16 percentage points more likely to enter the *Gymnasium*. This fact has potentially far-reaching implications for the treated children's probability of entering university and their labor market outcomes because children who complete the *Gymnasium* are much more likely to go to university and earn substantially higher salaries. Taken together, these findings suggest that the treated children derived substantial benefits from the WM training.

¹⁷ We estimate the costs of our intervention in a back-of-an-envelope calculation to be around US\$ 300 per child—which is considerably lower than the cost of the above-mentioned class size reduction by approximately 8 children. Our estimated costs include the cost for a software license (US\$ 20), the cost for notebooks or tablets of around US\$ 250 per child, and a budget for teacher training of US\$ 30 per child (assuming an intensive training session for teachers lasting for four hours costing US\$ 600 per teacher and 20 children per teacher as an average class-size). The cost of a WM training that becomes part of a general teaching routine and is at a similar scale of 25 school lessons, would be substantially lower as IT infrastructure and know-how would already be in place (and could be shared across classes).

References

- Ackerman, P. L., M. E. Beier, and M. O. Boyle. "Working Memory and Intelligence: The Same or Different Constructs?" *Psychological Bulletin* 131, no. 1 (2005): 30-60.
- Aksayli, N. D., G. Sala, and F. Gobet. "The Cognitive and Academic Benefits of Cogmed: A Meta-Analysis." *Educational Research Review* 27 (2019): 229-43.
- Alan, S., T. Boneva, and S. Ertac. "Ever Failed, Try Again, Succeed Better: Results from a Randomized Educational Intervention on Grit." *Quarterly Journal of Economics* 134, no. 3 (2019): 1121-62.
- Alan, S., and S. Ertac. "Fostering Patience in the Classroom: Results from Randomized Educational Intervention." *Journal of Political Economy* 126, no. 5 (2018): 1865-911.
- Alloway, T. P., and R. G. Alloway. "Investigating the Predictive Roles of Working Memory and Iq in Academic Attainment." *Journal of Experimental Child Psychology* 106, no. 1 (2010): 20-29.
- Almond, D., and J. Currie. "Human Capital Development before Age Five." *Handbook of Labor Economics, Vol 4b* 4 (2011): 1315-486.
- Almond, D., J. Currie, and V. Duque. "Childhood Circumstances and Adult Outcomes: Act Ii." *Journal of Economic Literature* 56, no. 4 (2018): 1360-446.
- Angrist, J. D., and V. Lavy. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *Quarterly Journal of Economics* 114, no. 2 (1999): 533-75.
- Au, J., E. Sheehan, N. Tsai, G. J. Duncan, M. Buschkuehl, and S. M. Jaeggi. "Improving Fluid Intelligence with Training on Working Memory: A Meta-Analysis." *Psychonomic Bulletin & Review* 22, no. 2 (2015): 366-77.
- Baddeley, Alan D. . *Essentials of Human Memory*. Hove, England: Psychology Press, 1999.
- Bell, Robert M., and Daniel F. McCaffrey. "Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples." *Survey Methodology* 28, no. 2 (2002): 169-81.
- Bellenberg, Gabriele. *Schulformwechsel in Deutschland*: Bertelsmann Stiftung, 2012.
- Borghans, L., A. L. Duckworth, J. J. Heckman, and B. ter Weel. "The Economics and Psychology of Personality Traits." *Journal of Human Resources* 43, no. 4 (2008): 972-1059.
- Bulheller, S., and H. O. Häcker. *Coloured Progressive Matrices (Cpm). Deutsche Bearbeitung Und Normierung Nach J. C. Raven*. Frankfurt: Pearson Assessment, 2010.
- Campbell, F., G. Conti, J. J. Heckman, S. H. Moon, R. Pinto, E. Pungello, and Y. Pan. "Early Childhood Investments Substantially Boost Adult Health." *Science* 343, no. 6178 (2014): 1478-85.
- Cappelen, Alexander W., John A. List, Anya Samek, and Beretil Tungodden. "The Effect of Early Education on Social Preferences." *Journal of Political Economy* (forthcoming).
- Carlana, M., E. La Ferrara, and P. Pinotti. "Goals and Gaps: Educational Careers of Immigrant Children." Hks Faculty Research Working Paper Series Rwp18-036, August 2018." *Harvard Kennedy School Faculty Research Working Paper Series RWP18-036* (2018).
- Carpenter, P. A., M. A. Just, and P. Shell. "What One Intelligence Test Measures - a Theoretical Account of the Processing in the Raven Progressive Matrices Test." *Psychological Review* 97, no. 3 (1990): 404-31.
- Constantinidis, C., and T. Klingberg. "The Neuroscience of Working Memory Capacity and Training." *Nature Reviews Neuroscience* 17, no. 7 (2016): 438-49.
- Conti, G., and J. J. Heckman. "Understanding the Early Origins of the Education-Health Gradient: A Framework That Can Also Be Applied to Analyze Gene-Environment Interactions." *Perspectives on Psychological Science* 5, no. 5 (2010): 585-605.
- Cunha, F., and J. Heckman. "The Technology of Skill Formation." *American Economic Review* 97, no. 2 (2007): 31-47.

- Cunha, F., and J. J. Heckman. "The Economics and Psychology of Inequality and Human Development." *Journal of the European Economic Association* 7, no. 2-3 (2009): 320-64.
- Cunha, F., J. J. Heckman, and S. M. Schennach. "Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Econometrica* 78, no. 3 (2010): 883-931.
- Cunha, Flavio, James Heckman, Lance Lochner, and Dimitriy Masterov. "Interpreting the Evidence on Life Cycle Skill Formation." In *Handbook of the Economics of Education*, edited by Erich a. Hanushek and Finis Welch, 698 - 812. Amsterdam: Elsevier B. V., 2006.
- de Abreu, Pmje, A. R. A. Conway, and S. E. Gathercole. "Working Memory and Fluid Intelligence in Young Children." *Intelligence* 38, no. 6 (2010): 552-61.
- Diamond, A., and K. Lee. "Interventions Shown to Aid Executive Function Development in Children 4 to 12 Years Old." *Science* 333, no. 6045 (2011): 959-64.
- Diamond, Adele, and Daphe S. Ling. "Review of the Evidence on, and Fundamental Questions About, Efforts to Improve Executive Functions, Including Working Memory." In *Cognitive and Working Memory Training*, edited by Jared M. Novick, Michael F. Bunting, Michael R. Dougherty and Randall Engle, W., 143 - 431. New York: Oxford University Press, 2020.
- Duckworth, A. L. "The Significance of Self-Control." *Proceedings of the National Academy of Sciences of the United States of America* 108, no. 7 (2011): 2639-40.
- Duckworth, A. L., and S. M. Carlson. "Self-Regulation and School Success." *Self-Regulation and Autonomy: Social and Developmental Dimensions of Human Conduct* (2013): 208-30.
- Duckworth, A. L., C. Peterson, M. D. Matthews, and D. R. Kelly. "Grit: Perseverance and Passion for Long-Term Goals." *Journal of Personality and Social Psychology* 92, no. 6 (2007): 1087-101.
- Duckworth, A. L., D. Weir, E. Tsukayama, and D. Kwok. "Who Does Well in Life? Conscientious Adults Excel in Both Objective and Subjective Success." *Frontiers in Psychology* 3 (2012).
- Dustmann, C. "Parental Background, Secondary School Track Choice, and Wages." *Oxford Economic Papers-New Series* 56, no. 2 (2004): 209-30.
- Dweck, C. S. *Minset: The New Psychology of Success*. New York: Random House, 2006.
- Eigsti, I. M., V. Zayas, W. Mischel, Y. Shoda, O. Ayduk, M. B. Dadlani, M. C. Davidson, J. L. Aber, and B. J. Casey. "Predicting Cognitive Control from Preschool to Late Adolescence and Young Adulthood." *Psychological Science* 17, no. 6 (2006): 478-84.
- Engle, R. W. "Working Memory Capacity as Executive Attention." *Current Directions in Psychological Science* 11, no. 1 (2002): 19-23.
- Esser, G., A. Wyszkon, and K. Ballaschk. *Basisdiagnostik Umschriebener Entwicklungsstörungen Im Grundschulalter (Buega)* Göttingen Hogrefe, 2008.
- Frison, L., and S. J. Pocock. "Repeated Measures in Clinical-Trials - Analysis Using Mean Summary Statistics and Its Implications for Design." *Statistics in Medicine* 11, no. 13 (1992): 1685-704.
- Gaspar, J. M., G. J. Christie, D. J. Prime, P. Jolicoeur, and J. J. McDonald. "Inability to Suppress Salient Distractors Predicts Low Visual Working Memory Capacity." *Proceedings of the National Academy of Sciences of the United States of America* 113, no. 13 (2016): 3693-98.
- Gathercole, S. E., S. J. Pickering, C. Knight, and Z. Stegmann. "Working Memory Skills and Educational Attainment: Evidence from National Curriculum Assessments at 7 and 14 Years of Age." *Applied Cognitive Psychology* 18, no. 1 (2004): 1-16.
- Gawrilow, C., and P. M. Gollwitzer. "Implementation Intentions Facilitate Response Inhibition in Children with Adhd." *Cognitive Therapy and Research* 32, no. 2 (2008): 261-80.
- Gertler, P., J. Heckman, R. Pinto, A. Zanolini, C. Vermeersch, S. Walker, S. M. Chang, and S. Grantham-McGregor. "Labor Market Returns to an Early Childhood Stimulation Intervention in Jamaica." *Science* 344, no. 6187 (2014): 998-1001.
- Heckman, J. J. "Skill Formation and the Economics of Investing in Disadvantaged Children." *Science* 312, no. 5782 (2006): 1900-02.

- Heckman, J. J., J. Stixrud, and S. Urzua. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *Journal of Labor Economics* 24, no. 3 (2006): 411-82.
- Heckman, J., R. Pinto, and P. Savelyev. "Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes." *American Economic Review* 103, no. 6 (2013): 2052-86.
- Helmers, K. F., S. N. Young, and R. O. Pihl. "Assessment of Measures of Impulsivity in Healthy Male-Volunteers." *Personality and Individual Differences* 19, no. 6 (1995): 927-35.
- Hofmann, W., T. Gschwendner, M. Friese, R. W. Wiers, and M. Schmitt. "Working Memory Capacity and Self-Regulatory Behavior: Toward an Individual Differences Perspective on Behavior Determination by Automatic Versus Controlled Processes." *Journal of Personality and Social Psychology* 95, no. 4 (2008): 962-77.
- Holmes, J., S. E. Gathercole, and D. L. Dunning. "Adaptive Training Leads to Sustained Enhancement of Poor Working Memory in Children." *Developmental Science* 12, no. 4 (2009): F9-F15.
- Jensen, Arthur Robert. *The G Factor: The Science of Mental Ability. Human Evolution, Behavior and Intelligence*. Westport, Conn Praeger, 1998.
- Karbach, J., and P. Verhaeghen. "Making Working Memory Work: A Meta-Analysis of Executive-Control and Working Memory Training in Older Adults." *Psychological Science* 25, no. 11 (2014): 2027-37.
- Katz, B., and P. Shah. "The Jury Is Still out on Working Memory Training." *Jama Pediatrics* 170, no. 9 (2016): 907-08.
- Kibby, M. Y., S. E. Lee, and S. M. Dyer. "Reading Performance Is Predicted by More Than Phonological Processing." *Frontiers in Psychology* 5 (2014).
- Klingberg, T. "Neural Basis of Cognitive Training and Development." *Current Opinion in Behavioral Sciences* 10 (2016): 97-101.
- Klingberg, T., E. Fernell, P. J. Olesen, M. Johnson, P. Gustafsson, K. Dahlstrom, C. G. Gillberg, H. Forsberg, and H. Westerberg. "Computerized Training of Working Memory in Children with Adhd - a Randomized, Controlled Trial." *Journal of the American Academy of Child and Adolescent Psychiatry* 44, no. 2 (2005): 177-86.
- Kosse, F., T. Deckers, P. Pinger, H. Schildberg-Horisch, and A. Falk. "The Formation of Prosociality: Causal Evidence on the Role of Social Environment." *Journal of Political Economy* 128, no. 2 (2020): 434-67.
- Krueger, A. B. "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics* 114, no. 2 (1999): 497-532.
- Kyllonen, P. C., and R. E. Christal. "Reasoning Ability Is (Little More Than) Working Memory Capacity." *Intelligence* 14, no. 4 (1990): 389-433.
- Martinussen, R., J. Hayden, S. Hogg-Johnson, and R. Tannock. "A Meta-Analysis of Working Memory Impairments in Children with Attention-Deficit/Hyperactivity Disorder." *Journal of the American Academy of Child and Adolescent Psychiatry* 44, no. 4 (2005): 377-84.
- McKenzie, D. "Beyond Baseline and Follow-Up: The Case for More T in Experiments." *Journal of Development Economics* 99, no. 2 (2012): 210-21.
- Melby-Lervag, M., and C. Hulme. "Is Working Memory Training Effective? A Meta-Analytic Review." *Developmental Psychology* 49, no. 2 (2013): 270-91.
- Melby-Lervag, M., T. S. Redick, and C. Hulme. "Working Memory Training Does Not Improve Performance on Measures of Intelligence or Other Measures of 'Far Transfer': Evidence from a Meta-Analytic Review." *Perspectives on Psychological Science* 11, no. 4 (2016): 512-34.
- Moffitt, T. E., L. Arseneault, D. Belsky, N. Dickson, R. J. Hancox, H. Harrington, R. Houts, R. Poulton, B. W. Roberts, S. Ross, M. R. Sears, W. M. Thomson, and A. Caspi. "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, no. 7 (2011): 2693-98.
- Nutley, S. B., and S. Soderqvist. "How Is Working Memory Training Likely to Influence Academic Performance? Current Evidence and Methodological Considerations." *Frontiers in Psychology* 8 (2017).
- Oberauer, K., R. Schulze, O. Wilhelm, and H. M. Suss. "Working Memory and Intelligence - Their Correlation and Their Relation: Comment on Ackerman, Beier, and Boyle (2005)." *Psychological Bulletin* 131, no. 1 (2005): 61-65.
- Redick, T. S., A. Calvo, C. E. Gay, and R. W. Engle. "Working Memory Capacity and Go/No-Go Task Performance: Selective Effects of Updating, Maintenance, and Inhibition." *Journal of Experimental Psychology-Learning Memory and Cognition* 37, no. 2 (2011): 308-24.
- Rhineland-Palatine, Statistic Office. *Allgemeinbildende Schulen Im Schuljahr 2017/2018*: Statistisches Landesamt Rheinland-Pfalz, 2018.
- Roberts, G., J. Quach, M. Spencer-Smith, P. J. Anderson, S. Gathercole, L. Gold, K. L. Sia, F. Mensah, F. Rickards, J. Ainley, and M. Wake. "Academic Outcomes 2 Years after Working Memory Training for Children with Low Working Memory a Randomized Clinical Trial." *Jama Pediatrics* 170, no. 5 (2016).
- Rode, C., R. Robson, A. Purviance, D. C. Geary, and U. Mayr. "Is Working Memory Training Effective? A Study in a School Setting." *Plos One* 9, no. 8 (2014).
- Romano, J. P., and M. Wolf. "Efficient Computation of Adjusted P-Values for Resampling-Based Stepdown Multiple Testing." *Statistics & Probability Letters* 113 (2016): 38-40.
- . "Stepwise Multiple Testing as Formalized Data Snooping." *Econometrica* 73, no. 4 (2005): 1237-82.
- Sala, G., N. D. Aksayli, K. S. Tatlidil, T. Tatsumi, Y. Gondo, and F. Gobet. "Near and Far Transfer in Cognitive Training: A Second-Order Meta-Analysis." *Collabra-Psychology* 5, no. 1 (2019).
- Sala, Giovanni, and Fernand Gobet. "Working Memory Training in Typically Developing Children: A Multilevel Meta-Analysis." *Psychonomic Bulletin & Review* <https://doi.org/10.3758/s13423-019-01681-y> (2020).
- Schmeichel, B. J., R. N. Volokhov, and H. A. Darnaree. "Working Memory Capacity and the Self-Regulation of Emotional Expression and Experience." *Journal of Personality and Social Psychology* 95, no. 6 (2008): 1526-40.
- Shipstead, Z., K. L. Hicks, and R. W. Engle. "Cogmed Working Memory Training: Does the Evidence Support the Claims?" *Journal of Applied Research in Memory and Cognition* 1, no. 3 (2012): 185-93.
- Sisk, V. F., A. P. Burgoyne, J. Z. Sun, J. L. Butler, and B. N. Macnamara. "To What Extent and under Which Circumstances Are Growth Mind-Sets Important to Academic Achievement? Two Meta-Analyses." *Psychological Science* 29, no. 4 (2018): 549-71.
- St Clair-Thompson, H., R. Stevens, A. Hunt, and E. Bolder. "Improving Children's Working Memory and Classroom Performance." *Educational Psychology* 30, no. 2 (2010): 203-19.
- Van Snellenberg, J. X., R. R. Girgis, G. Horga, E. van de Giessen, M. Slifstein, N. Ojeil, J. J. Weinstein, H. Moore, J. A. Lieberman, D. Shohamy, E. E. Smith, and A. Abi-Dargham. "Mechanisms of Working Memory Impairment in Schizophrenia." *Biological Psychiatry* 80, no. 8 (2016): 617-26.
- Westerberg, H., T. Hirvikoski, H. Forssberg, and T. Klingberg. "Visuo-Spatial Working Memory Span: A Sensitive Measure of Cognitive Deficits in Children with Adhd." *Child Neuropsychology* 10, no. 3 (2004): 155-61.
- Wiley, J., A. F. Jarosz, P. J. Cushen, and G. J. H. Colflesh. "New Rule Use Drives the Relation between Working Memory Capacity and Raven's Advanced Progressive Matrices." *Journal of Experimental Psychology-Learning Memory and Cognition* 37, no. 1 (2011): 256-63.

Yeager, D. S., P. Hanselman, G. M. Walton, J. S. Murray, R. Crosnoe, C. Muller, E. Tipton, B. Schneider, C. S. Hulleman, C. P. Hinojosa, D. Paunesku, C. Romero, K. Flint, A. Roberts, J. Trott, R. Iachan, J. Buontempo, S. M. Yang, C. M. Carvalho, P. R. Hahn, M. Gopalan, P. Mhatre, R. Ferguson, A. L. Duckworth, and C. S. Dweck. "A National Experiment Reveals Where a Growth Mindset Improves Achievement." *Nature* 573, no. 7774 (2019): 364-+.

Yeager, D. S., R. Johnson, B. J. Spitzer, K. H. Trzesniewski, J. Powers, and C. S. Dweck. "The Far-Reaching Effects of Believing People Can Change: Implicit Theories of Personality Shape Stress, Health, and Achievement During Adolescence." *Journal of Personality and Social Psychology* 106, no. 6 (2014): 867-84.

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 gratefully acknowledge financial support by the Jacobs Foundation (project 2013-1078-00), the German Academic Scholarship Foundation, the German Research Foundation (DFG), the university research priority program "Interdisciplinary Public Policy" at Johannes Gutenberg University Mainz, and the Research Council of Norway (FAIR, project 262675).

Online Appendix for
The Impact of Working Memory Training on Children's Cognitive and
Noncognitive Skills

Eva M. Berger¹, Ernst Fehr², Henning Hermes³, Daniel Schunk¹, Kirsten Winkel¹

¹Johannes Gutenberg University of Mainz, Department of Law and Economics, Jakob-Welder-Weg 4, 55128 Mainz, Germany. eva.berger@uni-mainz.de, daniel.schunk@uni-mainz.de, kirsten.winkel@uni-mainz.de

²University of Zurich, Department of Economics, Blümlisalpstrasse 10, 8006 Zurich, Switzerland. ernst.fehr@econ.uzh.ch

³Norwegian School of Economics, FAIR / Department of Economics, Helleveien 30, 5045 Bergen, Norway. henning.hermes@nhh.no

1	Supplementary Text.....	1
1.1	Supplementary Details on Participants.....	1
	Sampling of Participants.....	1
	Final Sample and Attrition.....	1
1.2	Supplementary Details on the Treatment.....	2
	Procedures.....	2
	Hardware.....	3
	Software.....	3
1.3	Supplementary Details on the Data Collection.....	4
	Testing Procedures.....	4
	Child, Parent, and Teacher Questionnaires.....	5
	Survey on Secondary School Track Choice.....	5
1.4	Supplementary Details on Outcome Measures.....	5
	Working Memory Tests (Figures S1–S3).....	5
	Educational Achievement Tests (Figures S4–S9).....	8
	Fluid IQ.....	14
	The Go/No-Go Task (Figure S10).....	14
	The bp Task (Figure S11).....	15
	Teacher-rated Self-regulation Skills.....	15
1.5	Supplementary Details on the Data Analysis.....	16
	Estimating the Treatment Effect.....	16
	Adjusting p-Values for Multiple Testing.....	17
	Decomposing the Treatment Effect of Working Memory Training.....	18
2	Further Supplementary Figures (Figures S12–S17).....	20
2.1	Distribution of Outcome Scores in W1–W4 (Figure S12–S15).....	20
2.2	Children’s Self-reported Motivation During Evaluation Tests (Figures S16–17).....	24
3	Supplementary Tables (Tables S1–S17).....	25
3.1	Tables Identifying the Treatment Effect.....	25
	Effects on Working Memory Capacity (Table S1).....	25
	Effects on Educational Outcomes and Fluid IQ (Table S2).....	26
	Effects on Performance in Go/No-Go Task and bp Task (Table S3).....	27
3.2	Heterogeneous Treatment Effects (Tables S4–S6).....	28
3.3	Controlling for Multiple Hypothesis Testing and Small Number of Clusters (Table S7).....	31
3.4	Controlling for Computer Use (Tables S8–S9).....	32
3.5	Tobit Estimates of Treatment Effects (Table S10).....	34
3.6	Restricting the Analyses to the No-attrition Sample (Tables S11–S13).....	35
3.7	Treatment Effects on Children’s Self-reported Motivation (Table S14).....	38
3.8	Factor Analysis of Teacher-rated Self-regulation Items (Table S15).....	39
3.9	Treatment Effects on Teacher-rated Self-regulation (Table S16).....	40
3.10	Decomposing the Treatment Effect of Working Memory Training (Table S17).....	41
4	References.....	42

1 Supplementary Text

The study was conducted in primary schools in Mainz, Germany in 2013/2014. It consisted of a five-week intervention and four data collection waves. We here provide supplementary details on participants (Section 1.1), the treatment condition (Section 1.2), the data collection waves (Section 1.3), outcome measures (Section 1.4), and the data analysis (Section 1.5). Supplementary figures are provided in Section 2, and all supplementary tables in Section 3.

The study consisted of a pre-intervention data collection wave (W1), the five-week intervention period, a data collection wave shortly after the intervention (W2), and two follow-up data collection waves after 6 and 12–13 months, respectively (W3 and W4).

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) one school lesson per day would be replaced by a working memory (WM) training lesson for 25 school days and (ii) the children would participate in all four planned data collection waves. In turn, schools received the IT infrastructure necessary to run the intervention, namely a notebook for each participating child (both for children assigned to the treatment as well as those assigned to the control group), rolling cases for transportation, charging and storage of the notebooks, as well as accessories like computer mice, headphones, and wifi routers. The schools retained this IT infrastructure for their permanent use.

Final Sample and Attrition

As described above, 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 W1–W4, resulting in a consent rate of 96.8%.¹ We were able to evaluate 572 children of the 580 for whom we received parental consent to collect data for our final data set². 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. Among the sample of 572 children, 292 were girls (51%) and 280 were boys (49%). Mean age at the beginning of the intervention (April 8, 2013) was 7.11 years (SD = 0.36 years).

¹ Among the children for whom we did not receive parental consent, roughly 50% participated in the working memory training while the other roughly 50% were in the control classes. However, we could not collect data in W1–W4 for these children. The participation of roughly half of these children in the working memory training without consent was possible because the school authorities viewed the training as part of regular teaching.

² Among the 572 children, 6 children participated in the baseline data collection (W1) somewhat after the start of the working memory training (because they were not available – due to illness – when the other children participated in this data collection). All reported effects of working memory training remain intact if we exclude these children from the data analysis.

Our sample decreased from 572 children in wave 1 (pre-training) to 531 children in wave 4 due to attrition. This corresponds to an attrition rate of 7.2%. Attrition did not differ between the treatment and control groups, the sample in the treatment group shrank from 279 to 259 children (attrition rate of 7.2%), while the sample in the control group shrank from 293 to 272 children (attrition rate of 7.2% as well). In previous WM training studies (see review paper by (Melby-Lervag and Hulme 2013)), the attrition rate was 10–11% even though the last follow-up measurements took place between three and eight months after the treatment in these studies. Thus, compared to these studies, our rate of attrition of roughly 7% over a period of more than a year is relatively low. Furthermore, 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 S11–13.

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 that of being assigned to the control group.

1.2 Supplementary Details on the Treatment

Procedures

The treatment in our study consisted of a daily WM training session that primarily took place during the first or second lesson at school over a period of 25 school days. The training was embedded into the classes' normal school routine. In each class, the teacher who covered the entire curriculum for the first grade also oversaw the study. The children thus considered the WM training to be a normal exercise unit, similar to when the teacher introduces new exercise units in a subject such as math, reading, or writing in the classroom. The teacher was present during the lessons when the WM training took place. The children also remained in their regular classroom and conducted the training sessions at their desks. This minimizes Hawthorne type effects because it ensures that the children viewed the WM training simply as a usual exercise unit in the context of their daily lessons, in which the sequential introduction of new learning content during the school year is part of normal school routine.

The first training session had an introductory character during which procedures and software were explained. The subsequent 24 lessons served as actual WM training sessions. The time frame for each training session was one school lesson, i.e. 50 minutes. During that time, every child had to pick up his/her computer as well as an external mouse and a headphone from the case, start the software, log-in, try to solve the training exercises, log-out, and put the notebook back to its pre-specified location. The net time available for training thus amounted to about 30 minutes per lesson.

The class teacher and one trained research assistant per class who helped the teacher (e.g. in distributing the notebooks, supporting the children during log-in, solving technical issues, ensuring compliance with the training protocol, and preparing a documentation of the training, including special events during training sessions) supervised the children.³ The assistants also helped in preparing a comprehensive documentation of the training.

³ The assistants were university students who were familiar with the working memory training software.

Hardware

Schools were equipped with one notebook for each child in the treatment *and* the control groups as well as large wheeled cases for storage, charging, and transportation of the notebooks. The cases also contained external mice and headphones for each child. For the treatment classes, each notebook was labeled with the child's name and his/her user account for the WM training software. The control group had no access to the WM training software.

Children only worked with the external mouse to ensure that the training group could not gain experience of any kind with an input device similar to the touchscreens used for the outcome measure tests in the data collection phases (see Section 1.3).

Software

The WM training software used for the treatment was “Cogmed RM”⁴ in an offline version with German instructions. It provides an age-specific user-interface, adaptive levels of difficulty, and a built-in incentive game (see below). The software requires the user to fulfill a certain set of tasks that consist of remembering sequences of information (e.g., numbers, locations) under various conditions. We excluded three of the thirteen different tasks available in the software because they contain letters or syllables that require reading abilities and knowledge about alphabetic characters that had not yet been introduced in all classes at the time of the WM training. Apart from this change (and the small reduction in trials, see below), we complied with the software provider’s required protocol.

Of the ten tasks implemented, two consisted of remembering spoken digits and, hence, focus on *verbal* WM capacity. These two tasks were very similar backward digit span tasks. The remaining eight tasks were based on remembering sequences of locations and visual information, and, thus, focused on *visuo-spatial* WM capacity. Due to the stronger emphasis on visuo-spatial relative to verbal WM training, we thus would expect larger improvements in visuo-spatial WM capacity.

Five of the ten training tasks were *simple span* tasks, as they only required storing and recalling information sequences of varying length. The remaining five tasks were *complex span* tasks because they contained at least one element of processing of stored content prior to recalling (e.g., numbers must be recalled in backward order or locations are moved before they have to be recalled).

The level of task difficulty was adapted based on the child’s previous performance. After a few correctly (incorrectly) solved trials, the level of difficulty increased (decreased). A daily training session consisted of six (varying) modules of 12 trials each (resulting in 72 trials per day)⁵. When the children had finished the six modules of a training session, they played a few trials of a fun game called “RoboRacing”. This is a feature built into the software and helps motivate children to participate in the WM training tasks.

Note that the training software was only available for the children during the five weeks of the intervention period. After this time, the login credentials for the software became invalid and no further training was thus possible. The software is, in principle, commercially available but was not so for the German market at the time of our intervention. Therefore, a further use of the training software after the time of our intervention was practically impossible (although the notebooks remained at the participating schools).

⁴ Cogmed and Cogmed Working Memory Training are trademarks, in the U.S. and/or other countries, of Pearson Education, Inc. or its affiliate(s).

⁵ The usual training protocol of Cogmed recommends 15 trials per module; we decreased the number of trials to 12 in order to fit the training in one school lesson (taking the time needed for picking up and bringing back the notebooks into account).

1.3 Supplementary Details on the Data Collection

The main data was collected at four points in time: wave 1 took place 3–4 weeks before the intervention (W1), wave 2 took place shortly after the intervention (W2), wave 3 took place 6 months after the intervention (W3), and wave 4 took place 12–13 months after the intervention (W4). In each wave, we collected several computer-based outcomes that served the purpose of measuring the consequences of WM training on skills. We describe these outcome measures in detail below. In addition, we administered questionnaires to teachers and parents. In W4, 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 8-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.

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.

Note that (a) our testing procedure guaranteed a high degree of standardization, especially through the instructions via headphones, and (b) by using large touchscreens as the method of data input, we ensured that there was no advantage for the treatment group as the computer-based WM training was run not with touchscreens but with a smaller notebook and external mice.

All tests were pretested in a 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.

Child Questionnaire

After the tests in W4, children were asked a few questions via headphone. Children answered on a 5-point Likert-type scale shown on the touchscreen. In particular, the questionnaire contained questions about how much effort the children put into completing the tests and how much they enjoyed the tasks: “How much did you enjoy doing the tasks on the computer just now?” and “How much did you try to do your best on the computer?” Figures S16 and S17 show the distribution of children’s answers for these questions. The great majority of children

reported that they provided high effort and had much fun in completing the tasks. Importantly, there are no differences between treatment and control group (see Table S14).

Parent Questionnaires

Parent questionnaires were only distributed in the data collection waves W1 and W3, i.e., before the intervention and 6 months after the intervention. Parent questionnaires included questions on socio-demographic characteristics of the family, parental behavior and characteristics as well as the child's attitude towards school and everyday behavior. Parents filled out 467 out of 572 parental questionnaires in W1 (82%) and 419 out of 544 in W3 (77%).

Teacher Questionnaires

In each data collection wave, teachers filled out a questionnaire containing questions on children's characteristics—such as their migration background or language problems—and teacher characteristics. In particular, we asked the teachers to assess each child's self-regulatory abilities using seven questions (see Section 1.4). Children's mean age on the day the teacher questionnaire was submitted equals to 85.3 months in W1, 88.0 months in W2, 92.8 months in W3, and 100.2 months in W4. We achieved a 100% return rate for the teacher questionnaire in all four evaluation waves, resulting in an $n = 572$ in W1 and W2, $n = 552$ in W3 and $n = 538$ in W4.

Survey on Secondary School Track Choice

In addition to the main data collection, we administered a short survey to parents when children were in the final grade of primary school (grade 4). This survey was conducted in April 2016, (i.e., about 3-4 years after the treatment) and asked parents about the secondary school track the child was about to enroll in from grade 5 on. 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. We received a total of 393 questionnaires (74.01% of the sample in W4). There was no difference in attrition between treatment and control group.

1.4 Supplementary Details on Outcome Measures

This section describes the tests that we used to measure the skill-consequences of WM training. WM capacity was assessed by one simple and two complex span tasks. For assessing educational achievement, we tested arithmetic skills, geometry skills, and reading comprehension. To measure important components of children's IQ, Raven's Coloured Progressive Matrices test (Raven 1995) was administered. For the assessment of self-regulation related abilities, we used a go/no-go task (adapted from (Gawrilow and Gollwitzer 2008) and the bp task (Esser, Wyszkon and Ballaschk 2008)). In a non-computer-based task we also measured children's time and risk preferences but these are not part of the current 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 the evaluation waves W1–W4 are displayed in Figures S12–S15.

Working Memory Tests

We adopted three different tasks for measuring the different facets of children's WM capacity. To avoid task-learning effects, we chose tasks distinct from the training tasks. The children's WM capacity was measured by a *verbal simple span* task, a *verbal complex span* task, and a *visuo-spatial complex span* task. The test scores in a given wave were constructed as follows.

We summed up the number of correctly solved item series weighted by each series' difficulty, which is defined by the series' length (i.e., number of items in the series).⁶ We standardized this score to mean = 0 and SD = 1. Because we expected the children to naturally improve their WM capacity when growing older, we increased the difficulty of the WM tasks across the four waves W1–W4 in order to avoid ceiling effects.

The *verbal simple span* task was a simple forward span short-term memory test. In this test, the child first had to listen to a sequence of one-digit numbers in the range of 1 to 9. After each sequence, a three by three grid with the digits 1 to 9 appeared on the screen and the child was asked to indicate the digits heard in the correct order (see Figure S1). The difficulty level in this task can be increased by increasing the number of items in the sequence of one-digit numbers that need to be recalled in the correct order.

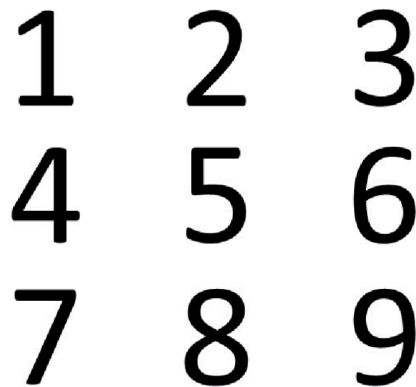


Figure S1: The Screen to Enter Answers for the Verbal Simple Span Task

In the *verbal complex span* task, the child first listened to a sequence of words, each of which described an object. After each object mentioned, the child had to decide whether the object is an animal or not by pushing a button “Animal” or “No animal”. Due to these “interruptions”, the task becomes a complex span WM task. After the sequence was finished, a three by three grid with pictures appeared on the screen. The pictures show the objects mentioned in the sequence as well as other, irrelevant objects. The child had to click on the pictures of the objects corresponding to the order in which the objects were previously mentioned (see Figure S2). The difficulty of this task was varied by varying the number of objects mentioned in a series.

⁶ We get the same results if we use the non-weighted sum of the correctly solved items series as a measure of working memory capacity.

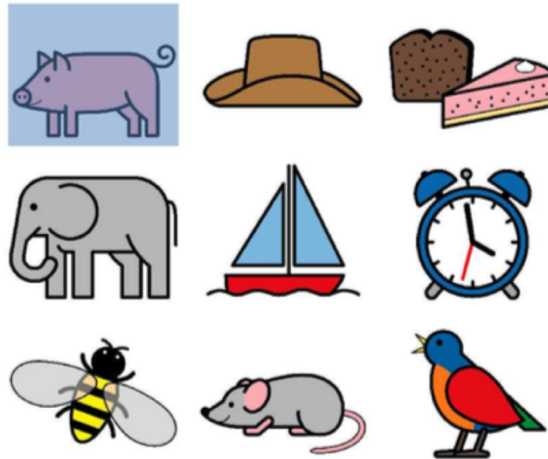


Figure S2: The Screen to Enter Answers for the Verbal Complex Span Task

The *visuo-spatial complex span* task was a complex span task measuring visuo-spatial WM capacity. First, the child was presented a sequence of “stimulus screens”. A stimulus screen contained three items; the child had to detect the item shaped differently and click on it (see Figure S3). Then, a new stimulus screen appeared and the child again had to click on the deviant shape, etc. Figure S3 below shows an example with three different stimulus screens after which the response screen appears which contains an empty grid. The child had to enter the position of the deviant items on the previous three stimulus screens in the correct order on the response screen. In Figure S3, for example, the correct response is to click “center”, “right”, “center” on the response screen. The difficulty level in this task is varied by varying the number of stimulus screens before the response screen appears.

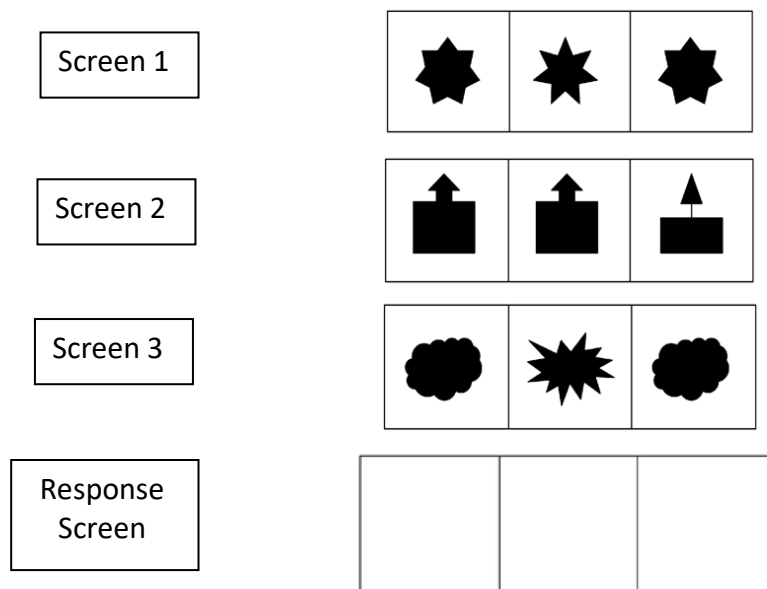


Figure S3: Stimulus and Response Screens in the Visuo-spatial Complex Span Task

Educational Achievement Tests

Educational achievement was assessed by testing for arithmetic skills, geometry skills, and reading skills. We increased the difficulty of the educational tasks across the four evaluation waves W1–W4 to avoid ceiling effects due to children’s development in scholastic skills with age.

Arithmetic skills: 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 Figure S4). 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 it 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.

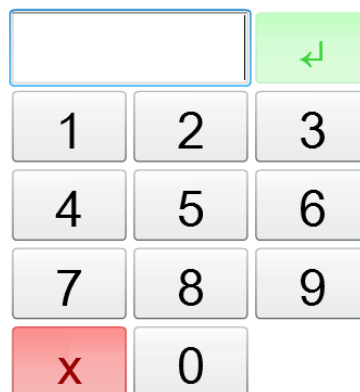


Figure S4: The Input Device for the Arithmetic and Geometry Tasks

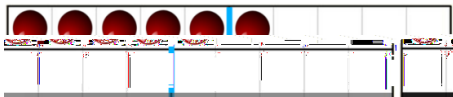
Number sense task: In this subtask, the children were presented a number of balls on a two by ten grid that was only shown for 1.7 seconds (see Figure S5 below showing several different examples with various levels of difficulty). In general, the display time was too short to count all balls before they 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 Figure S5) 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 Figure S5) 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.

Example for easy item:



Example for difficult item:

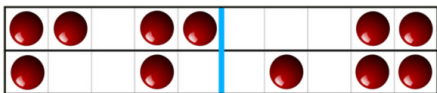
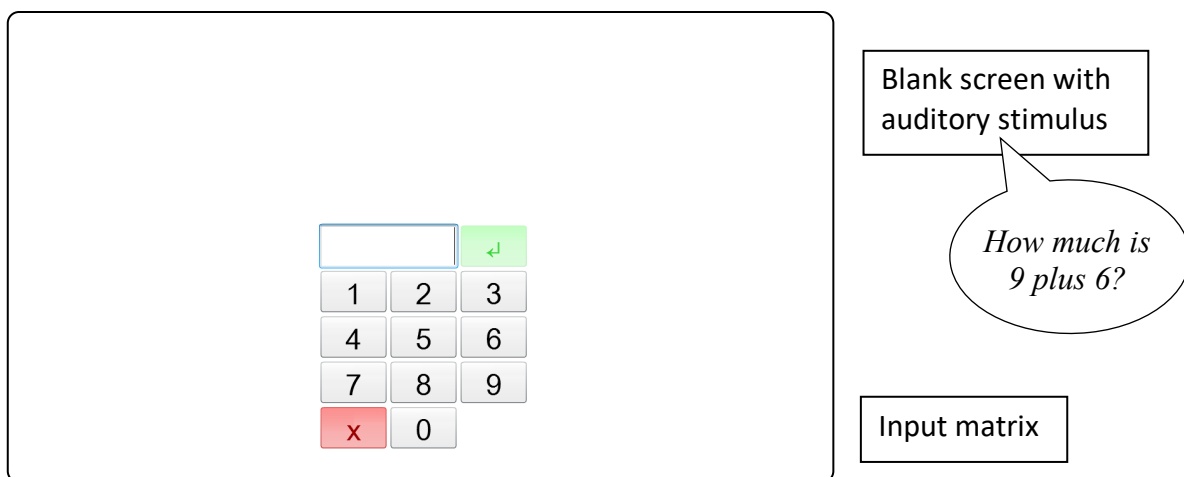


Figure S5: Number Sense Task, Screenshot Plus Two Further Examples

Auditory arithmetic task: This subtask measures arithmetic skills for addition and subtraction of two numbers (see Figure S6). Computational tasks were presented over the headphone (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 W1 and W2 the number range was up to 20, while in W3 and W4 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.



Example for easy item:

“How much is 2 plus 5?”

Example for difficult item:

“How much is 92 minus 17?”

Figure S6: Auditory Arithmetic Task, Screenshot Plus Two Further Examples

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 Figure S7). 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.

13 - 5 + 2 =

	↵	
1	2	3
4	5	6
7	8	9
x	0	

During the trial, the written stimuli were permanently visible

Example for easy item:

$$1 + 5 + 4 =$$

Example for difficult item:


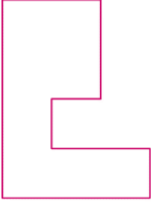
$$100 - 43 - 20 + 43 =$$

Figure S7: Written Arithmetic Task, Screenshot Plus Two Further Examples

The difficulty level was also adapted to the curriculum, analogously to the way it was done in the auditory arithmetic task.

Computation of final arithmetic test score: For each of the three subtasks (number sense, auditory and written arithmetic tasks), 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 three 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.


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 Figure S8 below). Depending on the size and the shape of the larger geometric object, this task can be made harder or easier.

How many  fit in  ?

1	2	3
4	5	6
7	8	9
x	0	

Stimulus and
Input matrix
permanently visible

Example for easy item:

How many  fit in  ?

Example for difficult item:

How many  fit in  ?

Figure S8: Geometry Task, Screenshot Plus Two Further Examples

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 Figure S8 (the red square): children could solve the problem without any mental 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 Figure S8 (the pink rectangle), children had to mentally rotate the small object to solve the question. For the difficult item in Figure S8 (the green triangle), children had 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.

Reading comprehension skills: Reading comprehension was assessed by a sentence comprehension test in single choice format. On the screen (see Figure S9), 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 tapping on the green enter-button right beside the sentence.

Nina does not want to go to school, to the zoo.

because

but

than

except

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 S9: Reading Comprehension Task, Screenshot Plus Two Further Examples

Generally, there was only one word missing in the sentence. In W3 and W4 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 W1 and W2, 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 W3, the reading comprehension task contained 16 sentences with 4 to 15 words per sentence in W3, and 16 sentences with 4 to 16 words per sentence in W4.

Fluid IQ

Children's fluid IQ was measured using a set of Colored Progressive Raven's Matrices (Raven 1995). While no single measurement tool will cover all aspects of a construct like fluid IQ, there is probably a broad consensus that the Raven's Matrices task captures important aspects of fluid IQ. We used two different sets of 17 items in W1/W3 and W2/W4, respectively. The child was shown a box with a pattern and had to choose which one out of six smaller patterns would fit into a missing part of the large pattern. The outcome score used in the main analysis is the standardized sum of correctly solved items.

The Go/No-Go Task

To measure inhibitory abilities, we employed a go/no-go task that was adapted from Gawrilow and Gollwitzer (Gawrilow and Gollwitzer 2008). 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 Figure S10 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 W2, W3, and W4.) 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 W1, W2, W3, and W4, respectively. In W1 and W3, the pictures were animals as described above. The pictures were vehicles in W2 and W4 (go = car, train, ship, airplane; no-go = truck).

We measure performance in this task in two ways. First, we simply 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 wave. Thus, a higher score indicates better performance in the task (i.e., fewer mistakes). Second, we compute the d' -measure of performance. The d' -measure is the standardized fraction of commission errors in the no-go items subtracted from the standardized fraction of correct responses in the go items. We again standardize this score to facilitate better interpretation. We find similar treatment effects with both performance measures, i.e., a significant increase in the performance of the treatment group relative to the control group in W4 (after in W2 and W3 the treatment group shows improved reaction times).

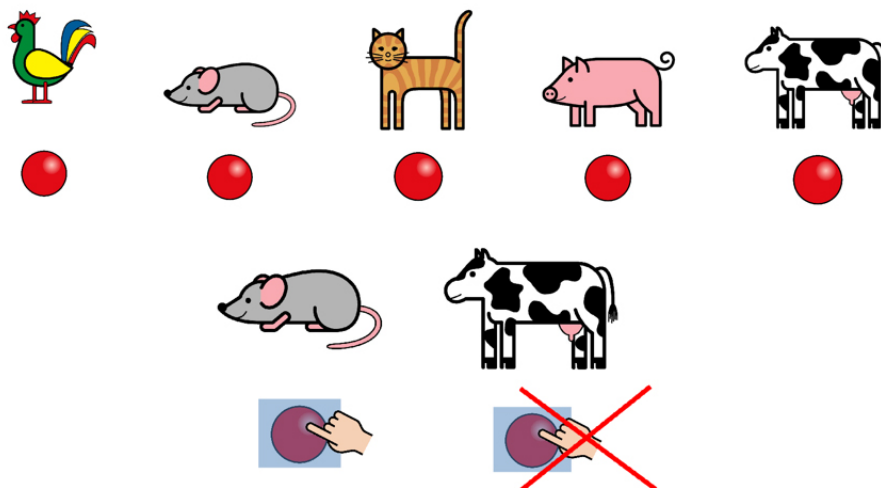


Figure S10: The Animals and the “Go-Button” in the Go/No-Go Task

The bp Task

The bp task measures *sustained concentration* and is taken from Esser et al. (Esser, Wyschkon and Ballaschk 2008). In this task, the child sees three lines filled with the letters “b”, “d”, “g”, “q”, “h”, and “p”, in total 45 letters on the touchscreen (see Figure S11 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” are displayed at the top of the screen in a salient form so that the child is 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 W1). 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).

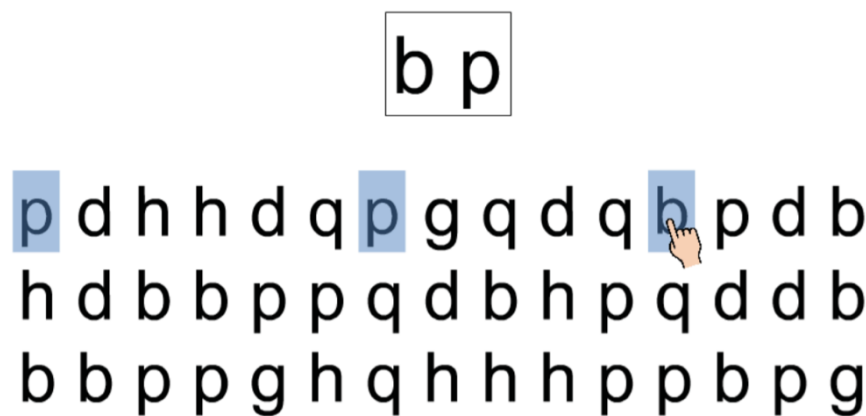


Figure S11: Example of a Screen in the bp Task

The bp task is an important control task in our field study that enables us to assess the specificity of the causal impact of WM training on other cognitive skills. WM capacity has been characterized as “executive attention”, i.e., an ability to maintain and implement a goal in the presence of distracting information (Engle 2002). If a task reinforces a given goal in every single trial—like in a specific type of stroop task in which almost all trials present incongruent stimuli or in the bp task—WM capacity hardly should play a role in task performance because goal maintenance is ensured by the task (Engle 2002). This contrasts with the go/no-go task in which the children are in the “go-mode” most of the time, as the goal of not pressing the “go-button” is only rarely reinforced. Therefore, the maintenance of the goal of not pressing in case of a rarely encountered stimulus and the subsequent inhibition of the pre-potent response is much harder in this task. Thus, based on the executive attention view of WM capacity (Engle 2002) we would predict that WM capacity plays a more important role in the go/no-go task compared to the bp task.

Teacher-rated Self-regulation Skills

All above-presented tasks for measuring the skill consequences of WM training are based on objective tests and not on subjective assessments. One of these tasks—the go/no-go task—provides a measure of the extent to which the children are able to inhibit pre-potent impulses. As the ability to inhibit these impulses is often viewed as a component of self-regulation or self-control, we decided to complement this measure with teachers’ assessments of the children’s self-regulation skills. We are aware of the fact that the teachers’ subjective

assessments may be a less reliable measure than objective tests. However, if WM training affects the objective test measure and the subjective measure in similar ways, our confidence in the reliability of the treatment effect is strengthened.

Therefore, the teachers assessed the self-regulation abilities of each child in their class by answering the questions listed below in each data collection wave. Questions 1–5 were answered by means of a 7-point Likert-type scale where 1 = “is not at all the case” and 7 = “is completely so”. The answer options for the questions 6 and 7 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)

The items above were developed with the purpose of best assessing young children’s self-regulation skills in a classroom context; the items are leaned to some extent on the Strengths and Difficulties Questionnaire (SDQ) proposed by Goodman (Goodman 1997) and on the Self-Control Scale developed by Tangney et al. (Tangney, Baumeister and Boone 2004) which was translated into German and validated by Bertrams and Dickhäuser (Bertrams and Dickhauser 2009).

We conducted a factor analysis of the items mentioned above (see Table S15). The results of this analysis show that all items exhibit considerable loadings on a single factor. All other factors have eigenvalues of less than 1. Table S15 shows this for the teachers’ answers in W1; the same results basically hold for all waves, however.

We used the standardized sum of the standardized answer values of the seven questionnaire items listed above as the dependent variable when estimating the effect of the WM training on teacher-assessed self-regulation abilities.

1.5 Supplementary Details on the Data Analysis

Estimating the Treatment Effect

To estimate the treatment effect of the WM training, we regress outcome scores measured after the training period on a treatment indicator. We also include school fixed effects (because randomization was conducted within schools) and some basic control variables (gender, age in months on January 1, 2013, and age in months at the relevant test days). Furthermore, we conducted other treatments (unrelated to WM training) in the same sample, with a randomly chosen part of the WM treatment group and a randomly chosen part of the control group. We control for the other treatments as well as for their interactions in all our estimations. Finally, in the estimation of each outcome score we also control for the pre-training baseline (W1) level of that score. This is done instead of using the difference-in-differences estimator. 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

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

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

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

where ρ is the autocorrelation of the outcome measures and n is the number of observations. Thus, the advantage of the method we use is that the variance of the estimate is smaller, i.e., the treatment effect is estimated with higher precision if $\rho \neq 1$ (Frison and Pocock 1992; McKenzie 2012).

Adjusting p-Values for Multiple Testing

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 (i.e., rejecting null hypotheses that should not be rejected, i.e., finding a significant effect where is none) increases with the number of hypotheses we test. This has to be corrected in order not to arrive at wrong conclusions.

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 true dependence structure to control the family-wise error rate (FWER, see (Romano and Wolf 2005))—a technique which is increasingly used for large-scale intervention studies (see, for example, (Cunha, Heckman and Schennach 2010; Campbell et al. 2014; Gertler et al. 2014)). Furthermore, we use a newly introduced efficient method to adjust p-values according to this stepdown algorithm (Romano and Wolf 2016).

Finally, we also combine this method of controlling the family-wise error rate (FWER) with the BRL (biased-reduced linearization) correction method. This method accounts for potential biases in estimation of standard errors when the number of clusters is relatively small (Bell and McCaffrey 2002).

We generate families of outcome measures by bundling outcomes in a natural way for the purpose of multiple hypothesis testing. The first family consists of our WM outcomes (verbal simple span, verbal complex span, visuo-spatial complex span). The second family consists of our educational outcome measures (arithmetic, geometry, reading). The third family consists of our measure for fluid IQ, Raven’s matrices. The fourth family is related to the hypotheses that follow from the executive attention approach and concerns the outcome measures from the go/no-go task and the bp task. We ran $M=5'000$ 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 and Wolf (Romano and Wolf 2005; Romano and Wolf 2016). The resulting p-values are reported in Table S7.

The Role of Working Memory Capacity for the Treatment Effect on Far-Transfer Outcomes

The key rationale behind our WM training intervention is that the training-induced increases in WM capacity will eventually enable the children to perform better in tasks that require WM capacity. Thus, it is natural to hypothesize that WM capacity is a mediator of the treatment effects of WM training on far-transfer outcomes. To gain insights into the quantitative importance of WM capacity as a mediator we conducted a mediation analysis.

The goal of this analysis is to decompose the total treatment effect of WM training on far-transfer outcomes into an effect that is due to increases in WM capacity and an effect that is due to other, unexplained factors. The total treatment effect is estimated by the equation

$$T_i^k = \delta_0^k + \delta_{WMT}^k WMT_i + \delta_X^k X_i + \varepsilon_i^k, \quad (3)$$

where T_i^k denotes the score of far-transfer outcome k of child i , WMT_i is the treatment indicator, and X_i denotes a vector of control variables such as age, gender, and the pre-treatment outcome score. The δ -parameters are to be estimated and ε_i^k is the error term. We carry out a decomposition analysis of the total treatment effect δ_{WMT}^k , focussing on those far-transfer outcomes that were significantly affected by the training: geometry, reading, Raven's IQ, and inhibitory ability in the Go/No-Go task.

Following Heckman, Pinto and Savelyev (2013), and similar applications by Kosse et al. (2020) and Carlana, La Ferrara and Pinotti (2018), our analysis is based on a linear production function for child i 's transfer outcome k , T_i^k . Thus, we assume that T_i^k is a function of working memory capacity, WMC_i , a vector of unknown mediating variables, U_i , and a vector of pre-program control variables, X_i :

$$T_i^k = \alpha_0^k + \alpha_{WMC}^k WMC_i + \alpha_U^k U_i + \alpha_X^k X_i + v_i^k \quad (4)$$

In equation (4), α_{WMC}^k is a parameter vector denoting the effect of WM capacity (verbal and visuo-spatial) on transfer outcome k ; α_U^k and α_X^k are parameter vectors related to the unknown mediating variables and pre-program control variables, respectively; v_i^k denotes an error term that is independent of the mechanisms and predetermined variables.⁷

In the first step of the decomposition analysis we estimate equation (4) based on the control group sample; the results are reported in Table S17 of this online appendix.⁸ The significantly positive estimates of α_{WMC}^k for all outcomes k suggests that working memory capacity plays a significant role in children's far-transfer outcomes.

In the second step of the decomposition analysis we estimate the treatment effect of the WM training on WM capacity, our mediation variable, by the following equation:

$$WMC_i = \beta_0 + \beta_{WMT} WMT_i + \beta_X X_i + \eta_i \quad (5)$$

The results of this regression are discussed in the results section of the paper (see Figure 1) and are reported in detail in Table S1.

In the third step, the decomposition is carried out in a straightforward way, assuming that program-induced increments in working memory capacity (WMC_i) and unmeasured mechanism variables (U_i) are statistically independent conditional on the controls X (following Heckman et al. (2013)). Taking the estimated parameters δ_{WMT}^k from equation (3) we

⁷ Note that the parameters in equation (4) are not indexed by the treatment indicator. This reflects the assumption that the parameters are the same for treatment and control group. This is consistent with the findings by Heckman et al. (2013), Carlana et al. (2018), as well as Kosse et al. (2020).

⁸ For the mediation analysis, we focus on those working memory capacity variables that are significantly affected by the WM training, i.e., the verbal simple span score and the visuo-spatial complex span score.

decompose it for each transfer outcome k into the part explained by working memory capacity improvements, $[(\alpha_{WMC}^k * \beta_{WMT}) / \delta_{WMT}^k] * 100\%$, and the unexplained part, $[1 - (\alpha_{WMC}^k * \beta_{WMT}) / \delta_{WMT}^k] * 100\%$.

2 Further Supplementary Figures S12–S14

2.1 Distribution of Outcome Scores in W1–W4

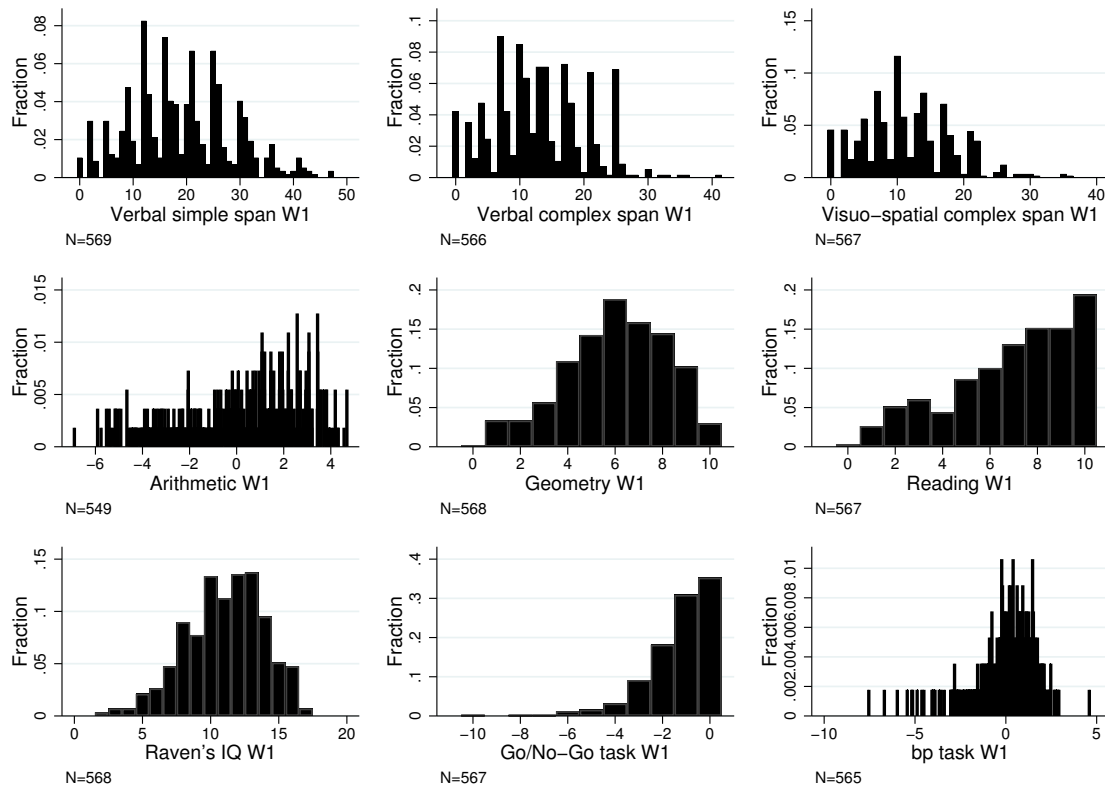


Figure S12: Distribution of Nonstandardized W1 Test Scores

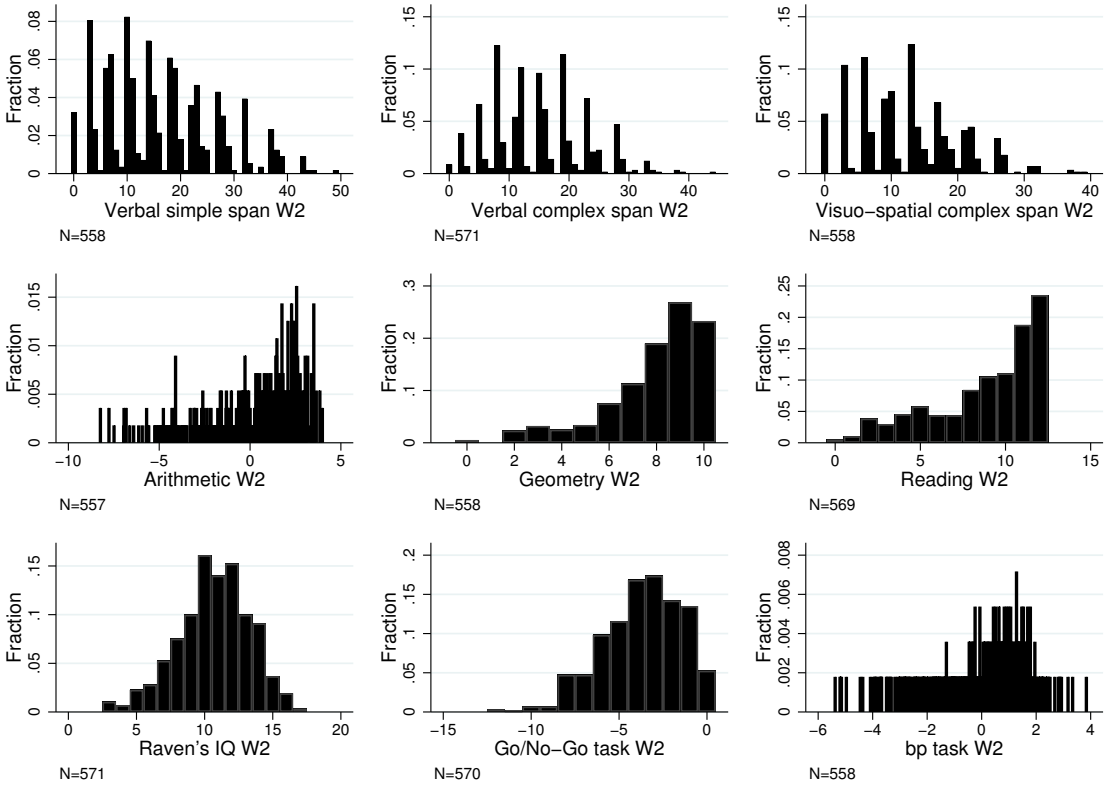


Figure S13: Distribution of Nonstandardized W2 Test Scores

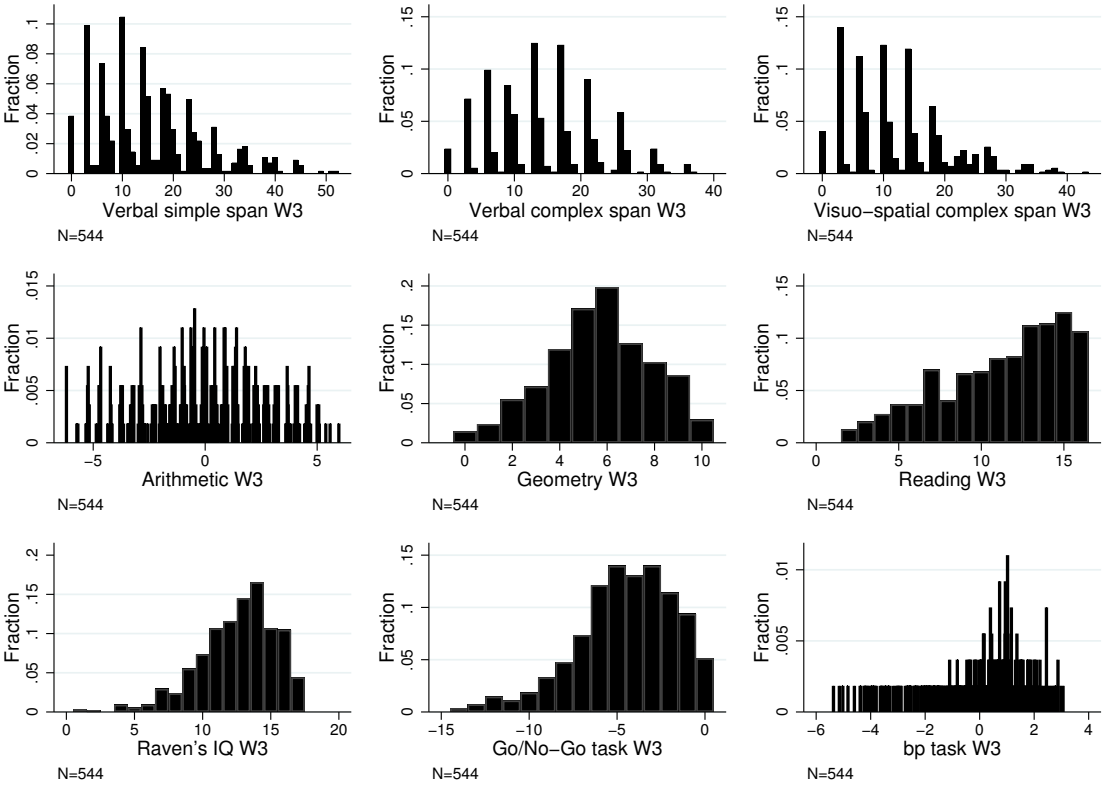


Figure S14: Distribution of Nonstandardized W3 Test Scores

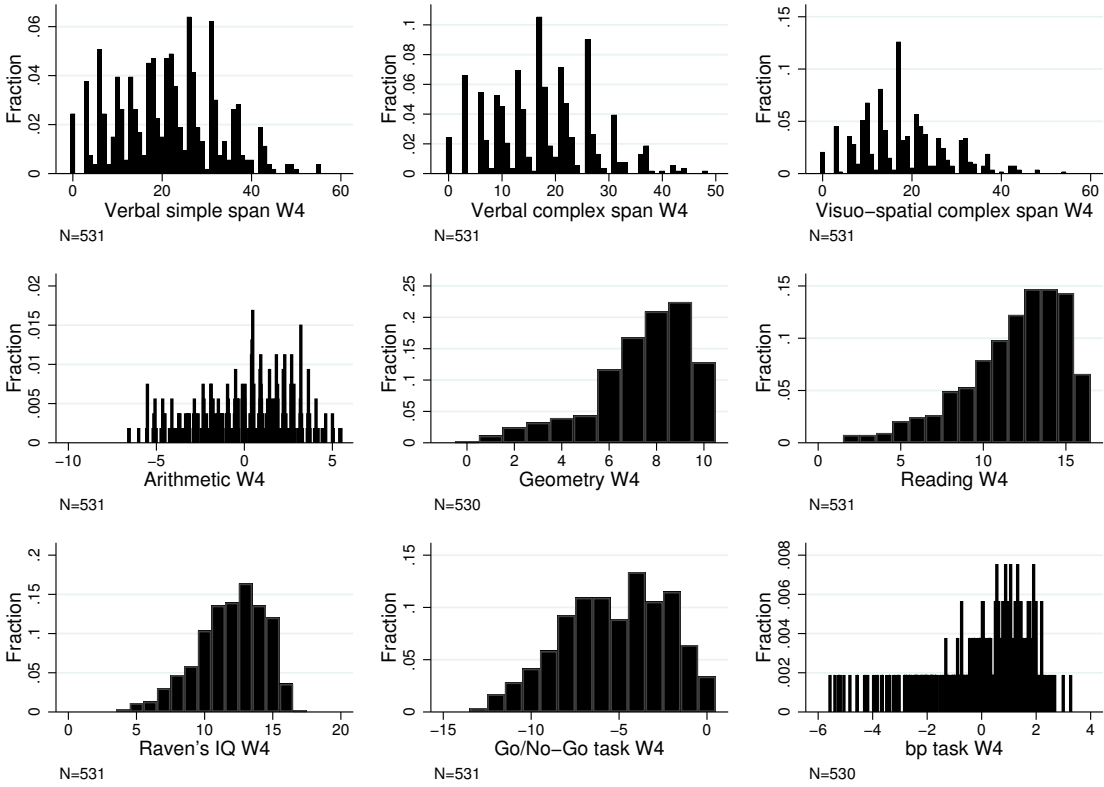
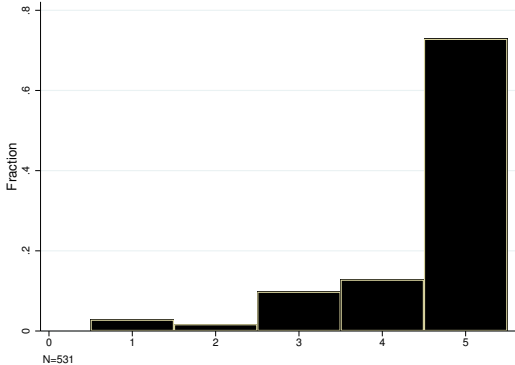


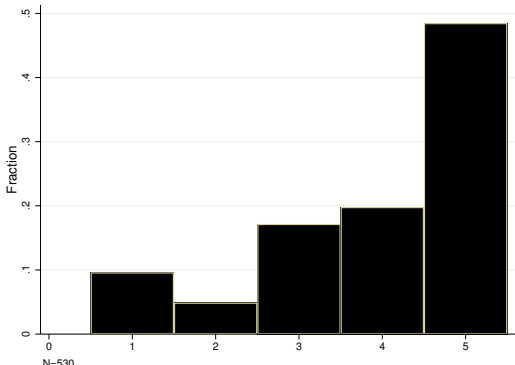
Figure S15: Distribution of Nonstandardized W4 Test Scores

2.2 Children’s Self-reported Motivation During Evaluation Tests



The figure plots children’s answers to the question “How much did you enjoy doing the tasks on the computer just now?” asked in a computerized short questionnaire immediately after the W4 evaluation. Answer options are a Likert-type scale ranging from 1 = “very little” to 5 = “very much”.

Figure S16: Children’s Enjoyment in W4 Tasks



The figure plots children’s answers to the question “How much did you try to do your best on the computer?” asked in a computerized short questionnaire immediately after the W4 evaluation. Answer options are a Likert-type scale ranging from 1 = “very little” to 5 = “very much”.

Figure S17: Children’s Effort in W4 Tasks

3 Supplementary Tables (Tables S1–S15)

3.1 Tables Identifying the Treatment Effect

Table S1: Treatment Effects on Working Memory Capacity

	Verbal simple span				Verbal complex span				Visuo-spatial complex span				
	W2 (1)	W3 (2)	W4 (3)	W4 (3)	W2 (4)	W3 (5)	W4 (6)	W4 (6)	W2 (7)	W3 (8)	W4 (9)	W3 (8)	W4 (9)
Working memory training	0.057 (0.058) [0.337]	0.382*** (0.062) [0.000]	0.295** (0.115) [0.015]	0.295** (0.115) [0.015]	-0.144 (0.108) [0.189]	-0.094 (0.059) [0.120]	0.032 (0.082) [0.702]	0.032 (0.082) [0.702]	0.395*** (0.105) [0.001]	0.458*** (0.078) [0.000]	0.443*** (0.108) [0.000]	0.458*** (0.078) [0.000]	0.443*** (0.108) [0.000]
N	555	541	528	528	565	539	526	526	554	540	527	540	527

The results are based on least squares models that regress the various standardized (mean = 0 and SD = 1) working memory scores on a dummy variable that takes on the value of 1 if the child received ‘working memory training’ and 0 otherwise. The regression also includes school fixed effects, the baseline outcome score (W1), and further controls (see Online Appendix Section 1.5 for details). W2, W3, and W4 refer to the evaluation waves shortly after, 6 months after, and 12–13 months after the working memory training period. Standard errors in parentheses are clustered at the classroom level. Related p-values are reported below in brackets (* p<0.10, ** p<0.05, *** p<0.01).

Table S2: Treatment Effects on Math, Reading, and Raven's IQ

	Arithmetic				Geometry				Reading				Raven's IQ				
	W2 (1)	W3 (2)	W4 (3)		W2 (4)	W3 (5)	W4 (6)		W2 (7)	W3 (8)	W4 (9)		W2 (10)	W3 (11)	W4 (12)		
Working memory training	-0.062 (0.092) [0.506]	-0.034 (0.081) [0.681]	-0.048 (0.086) [0.576]	512	0.166 (0.100) [0.108]	0.236** (0.097) [0.021]	0.384*** (0.101) [0.001]	527	0.022 (0.074) [0.765]	0.089 (0.099) [0.377]	0.230** (0.105) [0.037]	526	0.019 (0.066) [0.780]	0.237*** (0.075) [0.004]	540	0.237*** (0.070) [0.002]	527
N	535	525	512	554	541	527	564	567	539	539	526	567	540	527			

The results are based on least squares models that regress the various standardized (mean = 0 and SD = 1) outcome scores on a dummy variable that takes on the value of 1 if the child received 'working memory training' and 0 otherwise. The regression also includes school fixed effects, the baseline outcome score (W1), and further controls (see Online Appendix Section 1.5 for details). W2, W3, and W4 refer to the evaluation waves shortly after, 6 months after, and 12-13 months after the working memory training period. Standard errors in parentheses are clustered at the classroom level. Related p-values are reported below in brackets (* p<0.10, ** p<0.05, *** p<0.01).

Table S3: Treatment Effects on Performance in the Go/No-Go Task and bp Task

	Go/No-Go task				Go/No-Go: d'				bp task				
	W2 (1)	W3 (2)	W4 (3)	W4 (4)	W2 (4)	W3 (5)	W4 (6)	W4 (7)	W2 (7)	W3 (8)	W4 (9)	W2 (8)	W3 (9)
Working memory training	-0.088 (0.116) [0.454]	-0.021 (0.120) [0.862]	0.330*** (0.061) [0.000]	0.118 (0.142) [0.410]	0.118 (0.142) [0.410]	0.071 (0.142) [0.619]	0.475*** (0.084) [0.000]	0.049 (0.079) [0.539]	0.010 (0.084) [0.907]	0.073 (0.124) [0.559]	0.073 (0.124) [0.559]	0.010 (0.084) [0.907]	0.073 (0.124) [0.559]
N	566	540	527	565	565	540	527	552	538	524	538	524	524
R squared	0.134	0.142	0.120	0.137	0.137	0.180	0.116	0.344	0.245	0.212	0.245	0.212	0.212

The results are based on least squares models that regress the various standardized (mean = 0 and SD = 1) outcome scores on a dummy variable that takes on the value of 1 if the child received 'working memory training' and 0 otherwise. The regression also includes school fixed effects, the baseline outcome score (W1), and further controls (see Online Appendix Section 1.5 for details). W2, W3, and W4 refer to the evaluation waves shortly after, 6 months after, and 12-13 months after the working memory training period. Standard errors in parentheses are clustered at the classroom level. Related p-values are reported below in brackets (* p<0.10, ** p<0.05, *** p<0.01).

3.2 Heterogeneous Treatment Effects

Table S4: Heterogeneous Treatment Effects on Working Memory Capacity—Interaction with Low WM Capacity (W1)

	Verbal simple span				Verbal complex span				Visuo-spatial complex span			
	W2 (1)	W3 (2)	W4 (3)	W4 (3)	W2 (4)	W3 (5)	W4 (6)	W4 (6)	W2 (7)	W3 (8)	W4 (9)	W4 (9)
Working memory training	0.043 (0.067)	0.460*** (0.064)	0.350*** (0.108)	0.350*** (0.108)	-0.088 (0.117)	-0.045 (0.061)	0.099 (0.109)	0.099 (0.109)	0.354*** (0.107)	0.474*** (0.090)	0.441*** (0.116)	0.441*** (0.116)
Low WMC W1 (≤ 25 -perc.)	-0.311*** (0.077)	-0.090 (0.107)	-0.201 (0.136)	-0.201 (0.136)	-0.560*** (0.112)	-0.702*** (0.129)	-0.444*** (0.084)	-0.444*** (0.084)	-0.265*** (0.093)	-0.052 (0.095)	-0.326*** (0.109)	-0.326*** (0.109)
Low WMC W1 x WMT	0.119 (0.127)	-0.315** (0.149)	-0.319* (0.169)	-0.319* (0.169)	0.056 (0.157)	0.111 (0.157)	-0.072 (0.161)	-0.072 (0.161)	0.178 (0.123)	-0.073 (0.144)	-0.062 (0.154)	-0.062 (0.154)
N	549	535	522	522	560	535	522	522	549	535	522	522

This table examines whether the treatment effect of ‘working memory training’ on the standardized (mean = 0 and SD = 1) outcome scores for working memory capacity is different for children that are in the lowest quartile of working memory capacity at baseline. For this purpose, we use a dummy variable ‘Low WMC W1’ which takes on the value of 1 if working memory capacity in wave W1 is below or at the 25th percentile. The cardinal variable used for that is the sum of the three standardized working memory test scores. We interact ‘Low WMC W1’ with the treatment dummy. Apart from the inclusion of ‘Low WMC W1’ and its interaction with the treatment dummy, the least squares regressions also include school fixed effects, the respective baseline working memory scores, and further controls (see Online Appendix, Section 1.5). W2, W3, and W4 refer to the evaluation waves shortly after, 6 months after, and 12–13 months after the working memory training period. Standard errors in parentheses are clustered at the classroom level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. In this table, the coefficient related to the treatment dummy shows the treatment effect for children with a baseline working memory score above the 25th percentile. The results show that the treatment effects for these children are highly significant for verbal simple span in W3 and W4 and for visuo-spatial complex span in W2, W3, and W4. The coefficients related to the interaction term of ‘Low WMC W1’ with the treatment dummy show the extent to which the treatment effects are different for children with low working memory capacity at baseline. The table shows that—apart from the negative interaction effect for verbal simple span—there are no significant interaction effects. Also, for W4, the negative interaction for verbal simple span is no longer significant at the 5% level.

Table S5: Heterogeneous Treatment Effects on Math, Reading, and Raven's IQ—Interaction with Low WM Capacity (W1)

	Arithmetic				Geometry				Reading				Raven's IQ			
	W2 (1)	W3 (2)	W4 (3)	W4 (4)	W2 (4)	W3 (5)	W4 (6)	W4 (7)	W2 (7)	W3 (8)	W4 (9)	W2 (10)	W3 (11)	W4 (12)		
Working memory training	-0.052 (0.094)	0.018 (0.089)	-0.007 (0.088)	0.118 (0.109)	0.118 (0.109)	0.200** (0.093)	0.384*** (0.113)	0.088 (0.073)	0.088 (0.073)	0.132 (0.096)	0.317*** (0.094)	0.023 (0.068)	0.254*** (0.086)	0.263*** (0.091)		
Low WMC W1 (≤ 25 -perc.)	-0.257* (0.135)	-0.279** (0.111)	-0.334** (0.126)	-0.626*** (0.126)	-0.626*** (0.126)	-0.539*** (0.133)	-0.433** (0.205)	-0.224** (0.095)	-0.224** (0.095)	-0.358*** (0.112)	-0.254* (0.148)	-0.408*** (0.124)	-0.405*** (0.099)	-0.303** (0.123)		
Low WMC W1 x WMT	0.023 (0.179)	-0.131 (0.139)	-0.096 (0.156)	0.321** (0.153)	0.321** (0.153)	0.260 (0.207)	0.094 (0.257)	-0.157 (0.187)	-0.157 (0.187)	-0.034 (0.169)	-0.325 (0.215)	0.072 (0.139)	0.023 (0.135)	-0.175 (0.166)		
N	533	521	508	549	549	535	521	556	556	533	520	559	534	521		

This table examines whether the treatment effect of 'working memory training' on the standardized (mean = 0 and SD = 1) outcome scores for arithmetic, geometry, reading and Raven's IQ is different for children that are in the lowest quartile of working memory capacity at baseline. For this purpose, we use a dummy variable 'Low WMC W1' which takes on the value of 1 if working memory capacity in wave W1 is below or at the 25th percentile. The cardinal variable used for that is the sum of the three standardized working memory test scores. We interact 'Low WMC W1' with the treatment dummy. Apart from the inclusion of 'Low WMC W1' and its interaction with the treatment dummy, the least squares regressions also include school fixed effects, the respective baseline outcome scores, and further controls (see Online Appendix, Section 1.5). W2, W3, and W4 refer to the evaluation waves shortly after, 6 months after, and 12–13 months after the working memory training period. Standard errors in parentheses are clustered at the classroom level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. In this table, the coefficient related to the treatment dummy shows the treatment effect for children with a baseline working memory score above the 25th percentile. The results show that the treatment effects for these children are significant for geometry in W3 and W4, for reading in W4, and for Raven's IQ in W3 and W4. The coefficients related to the interaction term between 'Low WMC W1' and the treatment dummy show the extent to which the treatment effects are different for children with low working memory capacity at baseline. The table shows that—apart from the positive interaction effect for geometry in W2—there are no significant interaction effects.

Table S6: Heterogeneous Treatment Effects on Performance in the Go/No-Go Task and bp Task—Interaction with Low WM Capacity (W1)

	Go/No-Go task				bp task			
	W2 (1)	W3 (2)	W4 (3)	W4 (6)	W2 (4)	W3 (5)	W4 (6)	W4 (6)
Working memory training	-0.051 (0.136)	0.084 (0.092)	0.339*** (0.070)	0.166* (0.090)	0.068 (0.104)	0.068 (0.104)	0.117 (0.104)	0.117 (0.104)
Low WMC W1 (≤ 25 -perc.)	-0.265* (0.142)	-0.306** (0.120)	-0.357* (0.199)	0.075 (0.093)	-0.161 (0.134)	-0.161 (0.134)	-0.296 (0.200)	-0.296 (0.200)
Low WMC W1 x WMT	-0.058 (0.206)	-0.278 (0.254)	-0.019 (0.261)	-0.443*** (0.131)	-0.189 (0.185)	-0.189 (0.185)	-0.185 (0.236)	-0.185 (0.236)
N	558	533	520	547	532	532	518	518

This table examines whether the treatment effect of ‘working memory training’ on the standardized (mean = 0 and SD = 1) outcome scores for performance in the go/no-go task and the bp-task is different for children that are in the lowest quartile of working memory capacity at baseline. For this purpose, we use the dummy variable ‘Low WMC W1’ which takes on the value of 1 if working memory capacity in wave W1 is below or at the 25th percentile. The cardinal variable used for that is the sum of the three standardized working memory test scores. We interact ‘Low WMC W1’ with the treatment dummy. Apart from the inclusion of ‘Low WMC W1’ and its interaction with the treatment dummy, the least squares regressions also include school fixed effects, the respective baseline outcome scores, and further controls (see Online Appendix, Section 1.5). W2, W3, and W4 refer to the evaluation waves shortly after, 6 months after, and 12–13 months after the working memory training period. Standard errors in parentheses are clustered at the classroom level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. In this table, the coefficient related to the treatment dummy shows the treatment effect for children with a baseline working memory score above the 25th percentile. The results show that the treatment effect for these children is significant for performance in the go/no-go task in W4. The coefficient related to the interaction term between ‘Low WMC W1’ and the treatment dummy shows the extent to which the treatment effect is different for children with low working memory capacity at baseline. The table shows that—apart from the negative interaction effect for the bp-task in W2—there are no significant interaction effects.

3.3 Controlling for Multiple Hypothesis Testing and Small Number of Clusters

Table S7: Corrections for Multiple Hypotheses Testing and a Small Number of Clusters

	Wave 2	Wave 3	Wave 4
Panel A: Working Memory Capacity			
Verbal simple span	0.057 (0.697)	0.382*** (0.001)	0.295 (0.173)
Verbal complex span	-0.144 (0.561)	-0.094 (0.477)	0.032 (0.747)
Visuo-spatial complex span	0.395** (0.021)	0.458*** (0.001)	0.443** (0.021)
Panel B: Arithmetic, Geometry, and Reading			
Arithmetic	-0.062 (0.946)	-0.034 (0.946)	-0.048 (0.946)
Geometry	0.166 (0.543)	0.236 (0.228)	0.384** (0.025)
Reading	0.022 (0.946)	0.089 (0.893)	0.230 (0.302)
Panel C: Raven’s IQ			
Raven’s IQ	0.019 (0.809)	0.237** (0.041)	0.237** (0.041)
Panel D: Go/No-Go Task and bp Task			
Go/No-Go task	-0.088 (0.956)	-0.021 (0.981)	0.330*** (0.004)
bp task	0.049 (0.969)	0.010 (0.981)	0.073 (0.969)

The results are based on our main specifications reported in Supplementary Tables S1–S3. The coefficients are the point estimates showing how working memory training changes the outcome score indicated at the left-hand side of the table (as a fraction of a standard deviation) relative to the control group. We report p-values corrected for multiple hypothesis testing and small number of clusters in parentheses below each point estimate (* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$). The p-values are adjusted by controlling the family-wise error rate within each family of outcomes (corresponding to each panel in the table) using the step-down procedure by Romano & Wolf (2005, 2016), and by applying the conservative “biased reduced linearization (BRL) method” of Bell and McCaffrey (2002) to calculate clustered standard errors. The methods applied here are described in detail in Section 1.5 in the Online Appendix. Wave 2, Wave 3, and Wave 4 refer to the evaluation waves shortly after, 6 months after, and 12–13 months after the working memory training period. All treatment effects remain significant at the 5-percent level, except for the effect on verbal simple span in W4 (MHT-BRL corrected p-value = .173) and Reading in W4 (MHT-BRL corrected p-value = .302).

3.4 Controlling for Computer Use

Table S8: Treatment Effects on Working Memory Capacity—Controlling for Computer Use in Class

	Verbal simple span		Visuo-spatial comp.	
	W3 (1)	W4 (2)	W3 (3)	W4 (4)
Working memory training	0.352*** (0.051)	0.295** (0.118)	0.425*** (0.077)	0.444*** (0.110)
Use of computers in class W3	0.078* (0.040)		0.085* (0.049)	
Use of computers in class W4		0.004 (0.075)		-0.012 (0.051)
N	541	528	540	527

This table shows the estimates of the treatment effect of working memory training on the standardized (mean = 0 and SD = 1) outcome scores for verbal and visuo-spatial working memory when we additionally control for computer use in classes. The results are based on least squares models that regress the various working memory scores on a dummy variable that takes on the value of 1 if the child received ‘working memory training’ and 0 otherwise. The regression also includes school fixed effects, the baseline outcome score (W1), and further controls (see Online Appendix, Section 1.5). W3 and W4 refer to the evaluation waves 6 and 12–13 months after the working memory training period. The coefficients in the first row are point estimates showing how working memory training changes the working memory capacity scores 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Compared to Table S4, the coefficients in the first row of this table remain highly significant when we control for computer use.

Table S9: Treatment Effects on Math Geometry, Reading, Raven’s IQ, and Go/No-Go Task—Controlling for Computer Use in Class

	Geometry		Reading		Raven’s IQ		Go/No-Go task	
	W3 (1)	W4 (2)	W3 (3)	W4 (4)	W3 (5)	W4 (6)	W3 (7)	W4 (8)
Working memory training	0.215** (0.093)	0.377*** (0.096)	0.106 (0.099)	0.232** (0.105)	0.269*** (0.064)	0.234*** (0.069)	-0.045 (0.118)	0.329*** (0.063)
Use of computers in class W3	0.052 (0.057)		-0.043 (0.042)		-0.081** (0.035)		0.060 (0.049)	
Use of computers in class W4		0.055 (0.058)		-0.024 (0.061)		0.028 (0.048)		0.011 (0.042)
N	541	527	539	526	540	527	540	527

This table shows the estimates of the treatment effect of working memory training on the standardized (mean = 0 and SD = 1) outcome scores for geometry, reading, Raven’s IQ and performance in the go/no-go task when we additionally control for computer use in classes. The results are based on least squares models that regress the various outcome scores on a dummy variable that takes on the value of 1 if the child received ‘working memory training’ and 0 otherwise. The regression also includes school fixed effects, the baseline outcome score (W1), and further controls (see Online Appendix, Section 1.5). W3 and W4 refer to the evaluation waves 6 and 12–13 months after the working memory training period. The coefficients in the first row are point estimates showing how working memory training changes the outcome scores 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<0.10, ** p<0.05, *** p<0.01. Compared to Tables S5 and S6, the coefficients in the first row of this table remain highly significant when we control for computer use.

3.5 Tobit Estimates of Treatment Effects

Table S10: Treatment Effects on Performance in Math Geometry, Reading, and Go/No-Go Task—Tobit Estimation

	Geometry				Reading				Go/No-Go task			
	W2 (1)	W3 (2)	W4 (3)	W4 (3)	W2 (4)	W3 (5)	W3 (5)	W4 (6)	W2 (7)	W3 (8)	W3 (8)	W4 (9)
Working memory training	0.208* (0.122)	0.249** (0.098)	0.447*** (0.108)	0.447*** (0.108)	0.021 (0.093)	0.086 (0.110)	0.086 (0.110)	0.277*** (0.101)	-0.103 (0.115)	-0.012 (0.120)	-0.012 (0.120)	0.357*** (0.065)
N	554	541	527	527	564	539	539	526	566	540	540	527

The results are based on Tobit models that take the censoring of the dependent variable into account and regress the various standardized (mean = 0 and SD = 1) outcome scores on a dummy variable that takes on the value of 1 if the child received ‘working memory training’ and 0 otherwise. The regression also includes school fixed effects, the baseline outcome score (W1), and further controls (see Online Appendix, Section 1.5). W2, W3, and W4 refer to the evaluation waves shortly after, 6 months after, and 12–13 months after the working memory training period. Standard errors in parentheses are clustered at the classroom level. * p<0.10, ** p<0.05, *** p<0.01. The coefficients in the first row show the treatment effect (as a fraction of a standard deviation). It turns out that all coefficients that were significant in the least squares regression (see Tables S5 and S6) are also significant in the Tobit estimation.

3.6 Restricting the Analyses to the No-attrition Sample

Table S11: Treatment Effects on Working Memory Capacity—Sample Reduced to Children Appearing in All Waves

	Verbal simple span				Verbal complex span				Visuo-spatial complex span			
	W2 (1)	W3 (2)	W4 (3)	W4 (3)	W2 (4)	W3 (5)	W4 (6)	W4 (6)	W2 (7)	W3 (8)	W4 (9)	W4 (9)
Working memory training	0.089 (0.063)	0.394*** (0.046)	0.291*** (0.090)	0.291*** (0.090)	-0.131 (0.129)	-0.099 (0.058)	0.021 (0.080)	0.021 (0.080)	0.459*** (0.110)	0.454*** (0.068)	0.430*** (0.106)	0.430*** (0.106)
N	515	515	515	515	525	525	525	525	515	515	515	515

This table estimates the treatment effect of working memory training when we constrain the sample to those children that stay in the sample for all evaluation waves (from W1–W4). The results are based on least squares models that regress the various standardized (mean = 0 and SD = 1) working memory scores on a dummy variable that takes on the value of 1 if the child received ‘working memory training’ and 0 otherwise. The regression also includes school fixed effects, the baseline outcome score (W1), and further controls (see Online Appendix, Section 1.5). W2, W3, and W4 refer to the evaluation waves shortly after, 6 months after, and 12–13 months after the working memory training period. The coefficients in the first row are point estimates showing how working memory training changes the working memory 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<0.10, ** p<0.05, *** p<0.01.

Table S12: Treatment Effects on Math, Reading, and Raven's IQ—Sample Reduced to Children Appearing in All Waves

	Arithmetic				Geometry				Reading				Raven's IQ			
	W2 (1)	W3 (2)	W4 (3)		W2 (4)	W3 (5)	W4 (6)		W2 (7)	W3 (8)	W4 (9)		W2 (10)	W3 (11)	W4 (12)	
Working memory training	0.032 (0.099)	-0.034 (0.080)	-0.048 (0.077)	499	0.200** (0.095)	0.263*** (0.090)	0.356*** (0.101)	514	0.037 (0.073)	0.074 (0.098)	0.243** (0.102)	524	0.070 (0.076)	0.290*** (0.088)	0.248*** (0.072)	
N	499	499	499	499	514	514	514	514	524	524	524	524	526	526	526	

This table estimates the treatment effect of working memory training on various outcome scores when we constrain the sample to those children that stay in the sample for all evaluation waves (from W1–W4). The results are based on least squares models that regress the various standardized (mean = 0 and SD = 1) outcome scores on a dummy variable that takes on the value of 1 if the child received 'working memory training' and 0 otherwise. The regression also includes school fixed effects, the baseline outcome score (W1), and further controls (see Online Appendix, Section 1.5). W2, W3, and W4 refer to the evaluation waves shortly after, 6 months after, and 12–13 months after the working memory training period. The coefficients in the first row are point estimates showing how working memory training 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<0.10, ** p<0.05, *** p<0.01.

Table S13: Treatment Effects on Performance in the Go/No-Go Task and bp Task—Sample Reduced to Children Appearing in All Waves

	Go/No-Go task				bp task			
	W2 (1)	W3 (2)	W4 (3)	W4 (4)	W2 (4)	W3 (5)	W4 (6)	W4 (6)
Working memory training	-0.109 (0.121)	0.033 (0.109)	0.329*** (0.061)	0.103 (0.083)	0.002 (0.075)	0.062 (0.130)		
N	526	526	526	512	512	512	512	512

This table estimates the treatment effect of working memory training when we constrain the sample to those children that stay in the sample for all evaluation waves (from W1–W4). The results are based on least squares models that regress the various standardized (mean = 0 and SD = 1) outcome scores on a dummy variable that takes on the value of 1 if the child received ‘working memory training’ and 0 otherwise. The regression also includes school fixed effects, the baseline outcome score (W1), and further controls (see Online Appendix, Section 1.5). W2, W3, and W4 refer to the evaluation waves shortly after, 6 months after, and 12–13 months after the working memory training period. The coefficients in the first row are point estimates showing how working memory training 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<0.10, ** p<0.05, *** p<0.01.

3.7 Treatment Effects on Children’s Self-reported Motivation

Table S14: Treatment Effects on Children’s Self-Reported Motivation (W4)

	Enjoyment (W4) (1)	Effort (W4) (2)
Working memory training	-0.078 (0.104)	0.128 (0.156)
N	531	530

This table examines whether children who received ‘working memory training’ exert different effort in the evaluation tasks or enjoy them differently compared to children in the control group. The dependent variables used here are taken from the children’s answers to the two following questions that were asked immediately after completing the computer tasks 12–13 months after treatment: “How much did you enjoy doing the tasks on the computer just now?” (1 = “very little” to 5 = “very much”) and “How much did you try to do your best on the computer?” (1 = “very little” to 5 = “very much”). The results are based on least squares regressions including school fixed effects and further controls (see Online Appendix, Section 1.5). Standard errors in parentheses are clustered at the classroom level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3.8 Factor Analysis of Teacher-rated Self-regulation Items

Table S15: Factor Analysis of Teacher-rated Self-regulation Items (W1)

	Eigenvalue	
Factor 1	3.492459	
Factor 2	.7630324	
Factor 3	.0389734	
Factor 4	-.0677124	
Factor 5	-.0840027	
Factor 6	-.1307162	
Factor 7	-.134754	

	Factor 1	Factor 2
The child works in a concentrated and enduring manner.	-.8003659	.3931277
The child makes a large number of mistakes due to inattention.	.4946404	-.1795061
The child has a lot of self-discipline.	-.8409723	.2673741
The child has trouble waiting until it is his/her turn.	.6540294	.4583938
The child disturbs class instruction often.	.7583489	.3070454
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”)	.5795479	-.2831417
How do you rate the child with respect to patience? (1 = “very impatient” up to 7 = “very patient”)	-.7491641	-.3466996

This table shows the factor analysis for the items that were used to construct the teacher-rated self-regulation score W1. Questions 1–5 were answered by means of a 7-point Likert-type scale where 1 = “is not at all the case” and 7 = “is completely so”. The answer options for the questions 6 and 7 are indicated behind the question. The factor analysis shows that it is appropriate to extract only one factor. This factor has an eigenvalue > 1 and all survey items display strong factor loadings on this factor, while for all other factors the eigenvalue is clearly below one and factor loadings are small.

3.9 Treatment Effects on Teacher-rated Self-regulation

Table S16: Treatment Effects on Teacher-rated Self-regulation

	W2 (1)	W3 (2)	W4 (3)
Working memory training	0.224* (0.111)	0.369** (0.171)	0.269** (0.114)
N	555	527	517

This table reports the results of least squares regressions of the standardized (mean = 0 and SD = 1) teacher-rated self-regulation scores in the different evaluation waves on a dummy variable that takes on the value of 1 if the child received ‘working memory training’ and 0 otherwise. The regression also includes school fixed effects, the baseline outcome score (W1), and further controls (see Online Appendix, Section 1.5). W2, W3, and W4 refer to the evaluation waves shortly after, 6 months after, and 12–13 months after the working memory training period. The coefficients in the first row are the point estimates showing how working memory training changes the teacher-rated self-regulations scores in the various evaluation waves (as a fraction of a standard deviation) relative to the control group. Standard errors in parentheses are clustered at the classroom level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3.10 Decomposing the Treatment Effect of Working Memory Training

Table S17: The Relevance of Working Memory Capacity in Geometry, Raven's IQ, and Go/No-Go Task

	Geometry W4 (1)	Raven's IQ W4 (2)	Go/No-Go task W4 (3)
Visuo-spatial complex span W4	0.332*** (0.078) [0.001]	0.226*** (0.056) [0.001]	-0.015 (0.052) [0.783]
Verbal simple span W4	0.208*** (0.068) [0.008]	0.188*** (0.041) [0.000]	0.211*** (0.052) [0.001]
N	271	269	269

The results are based on least squares models that regress the various standardized (mean = 0 and SD = 1) far-transfer outcome scores from W4 (i.e., evaluation wave 12–13 months after the working memory training period) on the standardized outcomes scores for various working memory scores W4. All models additionally include school fixed effects, the pre-training baseline (W1) level of the respective outcome score, gender, age, and age at test day. The sample is restricted to the control group. Standard errors in parentheses are clustered at the classroom level. Related p-values are reported in brackets. * p<0.10, ** p<0.05, *** p<0.01.

4 References

- Bell, R.M. and D.F. McCaffrey. "Bias Reduction in Standard Errors for Linear Regression with Multistage Samples", *Survey Methodology* 28 (2002): 169-181.
- Bertrams, A., and O. Dickhauser. "Measuring Dispositional Self-Control Capacity. A German Adaptation of the Short Form of the Self-Control Scale (Scs-K-D)." *Diagnostica* 55, no. 1 (2009): 2-10.
- Campbell, F., G. Conti, J. J. Heckman, S. H. Moon, R. Pinto, E. Pungello, and Y. Pan. "Early Childhood Investments Substantially Boost Adult Health." *Science* 343, no. 6178 (2014): 1478-85.
- Carlana, M., E. La Ferrara, and P. Pinotti. "Goals and Gaps: Educational Careers of Immigrant Children." Hks Faculty Research Working Paper Series Rwp18-036, August 2018." *Harvard Kennedy School Faculty Research Working Paper Series RWP18-036* (2018).
- Cunha, F., J. J. Heckman, and S. M. Schennach. "Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Econometrica* 78, no. 3 (2010): 883-931.
- Engle, R. W. "Working Memory Capacity as Executive Attention." *Current Directions in Psychological Science* 11, no. 1 (2002): 19-23.
- Esser, G., A. Wyschkon, and K. Ballaschk. *Basisdiagnostik Umschriebener Entwicklungsstörungen Im Grundschulalter (Buega)* Göttingen Hogrefe, 2008.
- Frison, L., and S. J. Pocock. "Repeated Measures in Clinical-Trials - Analysis Using Mean Summary Statistics and Its Implications for Design." *Statistics in Medicine* 11, no. 13 (1992): 1685-704.
- Gawrilow, C., and P. M. Gollwitzer. "Implementation Intentions Facilitate Response Inhibition in Children with Adhd." *Cognitive Therapy and Research* 32, no. 2 (2008): 261-80.
- Gertler, P., J. Heckman, R. Pinto, A. Zanolini, C. Vermeersch, S. Walker, S. M. Chang, and S. Grantham-McGregor. "Labor Market Returns to an Early Childhood Stimulation Intervention in Jamaica." *Science* 344, no. 6187 (2014): 998-1001.
- Goodman, R. "The Strengths and Difficulties Questionnaire: A Research Note." *Journal of Child Psychology and Psychiatry and Allied Disciplines* 38, no. 5 (1997): 581-86.
- Heckman, J., R. Pinto, and P. Savelyev. "Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes." *American Economic Review* 103, no. 6 (2013): 2052-86.
- Kosse, F., T. Deckers, P. Pinger, H. Schildberg-Horisch, and A. Falk. "The Formation of Prosociality: Causal Evidence on the Role of Social Environment." *Journal of Political Economy* 128, no. 2 (2020): 434-67.
- McKenzie, D. "Beyond Baseline and Follow-Up: The Case for More T in Experiments." *Journal of Development Economics* 99, no. 2 (2012): 210-21.
- Melby-Lervag, M., and C. Hulme. "Is Working Memory Training Effective? A Meta-Analytic Review." *Developmental Psychology* 49, no. 2 (2013): 270-91.
- Raven, J.C. *Coloured Progressive Matrices*. Oxford, United Kingdom: Oxford Psychologists Press, 1995.
- Romano, J. P., and M. Wolf. "Efficient Computation of Adjusted P-Values for Resampling-Based Stepdown Multiple Testing." *Statistics & Probability Letters* 113 (2016): 38-40.
- . "Stepwise Multiple Testing as Formalized Data Snooping." *Econometrica* 73, no. 4 (2005): 1237-82.
- Tangney, J. P., R. F. Baumeister, and A. L. Boone. "High Self-Control Predicts Good Adjustment, Less Pathology, Better Grades, and Interpersonal Success." *Journal of Personality* 72, no. 2 (2004): 271-324.