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

Eva M. Berger

Federal Ministry of Labour and Social Affairs (Bundesministerium für Arbeit und Soziales)

Ernst Fehr

University of Zurich

Henning Hermes

ifo Institute Munich, Centre for Economic Policy Research, CESifo, FAIR (Centre for Experimental Research on Fairness, Inequality and Rationality), Human Capital and Economic Opportunity Global Working Group, and Institute of Labor Economics (IZA)

Daniel Schunk

Johannes Gutenberg University Mainz

Kirsten Winkel

University of Koblenz

Working-memory (WM) capacity is a key component of a wide range of cognitive and noncognitive skills—such as fluid IQ, math, reading, and inhibitory control—but can WM training improve these skills? Here, we

We 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 thank Michael Wolf for

Electronically published December 13, 2024

Journal of Political Economy, volume 133, number 2, February 2025.

© 2024 The University of Chicago. This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0), which permits non-commercial reuse of the work with attribution. For commercial use, contact journalpermissions@press.uchicago.edu. Published by The University of Chicago Press.

https://doi.org/10.1086/732884

examine the causal impact of WM training embedded in regular school teaching, using a randomized educational intervention with 6–7-year-old children. We find substantial gains in WM capacity and document positive spillover effects on geometry, fluid IQ, and inhibitory control. Three years later, treated children are 16 percentage points more likely to enter an advanced secondary school track.

I. Introduction

Cognitive and noncognitive 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). Executive functions (EFs; Diamond 2013), which are malleable through interventions in childhood with long-lasting effects into adulthood (see, e.g., Walker et al. 2022; García, Heckman, and Ronda 2023), are thought to play a key role in a wide range of abilities. Working-memory (WM) capacity—the ability to mentally store and process information (Baddeley 1999)—is a key component of EFs and has been shown to be 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; Engel 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 Dernaree 2008; Diamond and Ling 2020). Conversely, individuals with learning problems and 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 this relevance of EFs and WM capacity for many important skills, the question is whether one

support and provision of code in conducting the multiple testing correction. Further, we are grateful to three anonymous referees, and we thank the editor for pointing us to the multiple links of our study to research on executive functions and for making the paper more concise. We gratefully acknowledge financial support by the Jacobs Foundation (project 2013-1078-00), the University Research Priority Program of the University of Zurich on Equality of Opportunity (project U-302-01-01), the German Academic Scholarship Foundation, the German Research Foundation (DFG, BE 5436/1-1), the university research priority program “Interdisciplinary Public Policy” at Johannes Gutenberg University Mainz (project FI 2/2014-2016), and the Research Council of Norway (FAIR, project 262675). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Fehr and Schunk initiated and supervised the study throughout all stages. Fehr, Schunk, Berger, and Winkel conceptualized the study, and all the authors developed the field design. Berger, Hermes, Schunk, and Winkel developed outcome measures for the study. Hermes conducted the field experiment, with input from Berger, Fehr, Schunk, and Winkel; Berger and Hermes performed the data analysis, with input from Fehr, Schunk, and Winkel; all the authors were involved in the interpretation of the results, and all wrote the paper. This paper was edited by James J. Heckman.

can simultaneously improve several of these skills through WM training and whether this can be achieved by introducing WM training into the school 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 (direct effects). However, the question of whether training-induced improvements in WM capacity lead to improvements in other important skills, such as academic and self-regulatory skills (spillover 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-Lervåg, Redick, and Hulme 2016; Aksayli, Sala, and Gobet 2019; Sala et al. 2019; Gobet and Sala 2023). This lack of a conclusive answer suggests that WM-training studies face a number of considerable challenges (see, e.g., Greene et al. 2019 and Gobet and Sala 2023). For example, (i) potential spillover 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; and (iii) training may lead to spillover effects only 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 economics indicates that training programs for youths in their late adolescence or young adulthood may be less effective than those for young children (Cunha et al. 2006; Heckman 2006). Young children have higher brain plasticity, which might increase the chances of generating positive spillover effects (Heckman 2006; Constantinidis and Klingberg 2016; Klingberg 2016; Almond, Currie, and Duque 2018). In contrast to most other WM-training studies in typically developing children, we track children's outcomes for longer than 3–5 months after the training. Specifically, we also measure outcomes after 6 and after 12–13 months, and we examine whether the training has an effect on children's school trajectory 3 years later.

In our study, 31 school classes were randomly assigned to a treatment group (15 classes) or a control group (16 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 5 weeks. Not only do we find substantial direct effects on WM capacity that emerge right after the 5-week training period and last throughout all evaluation waves, we also find spillover effects on several important skills, such as geometry, Raven's fluid-IQ measure, and children's ability to inhibit prepotent impulses. Interestingly, for all these abilities there is no significant treatment effect shortly after the training; that is, the spillover 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 SD (standard deviations). These effects are sizeable in view of the intervention's intensity (25 school hours).

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 5-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 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 fourth grade in primary school is one of the most decisive educational choices for a child. This decision typically has a large influence on the probability of earning a high school (i.e., *Gymnasium*) degree and thus on later university enrollment and adult labor market outcomes.[1]

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,

[1] Dustmann (2004) finds that individuals with a degree from 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 ninth grade) and 22%–34% higher wages than individuals with an intermediate degree (*Realschulabschluss*, earned after tenth grade).

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 discussed in Borghans et al. (2008), Cunha and Heckman (2009), Almond, Currie, and Duque (2018), García and Heckman (2023), and García, Heckman, and Ronda (2023) 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 designed interventions to boost cognitive and noncognitive skills and have conducted randomized controlled trials to measure the interventions' causal effects. This literature examined, among others, the general role and malleability of (i) children's "growth mindset," that is, an optimistic belief about the role of effort in individuals' success (Dweck 2006; Yeager et al. 2014, 2019; Sisk et al. 2018), (ii) children's perseverance and patience (Duckworth et al. 2007; Duckworth 2011; Alan and Ertac 2018), and (iii) children's trust and social preferences (Cappelen et al. 2020; Kosse et al. 2020).

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: WM capacity. WM capacity is a key component of executive functioning (Diamond 2013)—with inhibitory control and cognitive flexibility being the other two components—and comprises not just the ability to store information in the short term. The use of WM also requires the ability to process information in the presence of distracting impulses and competing information that is not conducive for the individual's goal. Research on executive functioning has therefore emphasized that WM and inhibitory control "generally need one another and co-occur" and that WM "supports inhibitory control" (Diamond 2013, 143). This is the reason why WM capacity may also generate spillover effects on noncognitive skills by facilitating impulse control and self-regulation.

The literature on WM training in typically developing children has mostly measured the impact of WM training only immediately after the training or a few weeks or months after the training. There are, however, reasons to believe that detecting spillover effects to more complex skills might require follow-up evaluations that leave more time for spillover effects to develop. Cunha and Heckman (2007; also 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 as a result of "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 investment into skills at subsequent stages).[2] 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 spillover effects need time to develop: in all cases in which we eventually document a significant spillover effect, the effect is rising over time, but in none of these cases is the spillover effect already significant shortly after the training. However, after 6 months a spillover effect on geometry skills and Raven's fluid-IQ measure emerges (also visible after 12–13 months), and after 12–13 months we observe, in addition, a spillover effect on inhibitory control, namely, the ability to inhibit prepotent impulses.

Our paper is also related to the literature in psychology and education science that examines whether EF- and WM-training interventions (and other forms of cognitive training) lead to spillover effects in children (for an early contribution, see Klingberg et al. 2005; for reviews, see Diamond 2013, Diamond and Ling 2020, and Sala and Gobet 2020). A relevant share of this literature focuses on disadvantaged children, for example, with disorders or very low WM capacity or from low-education family backgrounds (e.g., Klingberg et al. 2005; Roberts et al. 2016). For interventions targeting these disadvantaged children, several studies show strong positive long-term effects on EFs that also spill over to several other domains, such as health, education, and (reduced) crime (Walker et al. 2022; García and Heckman 2023; García, Heckman, and Ronda 2023). Our paper instead focuses on typically developing children. We contribute to this literature by demonstrating positive WM-training effects, showing that improvements in one EF domain (WM) can create spillovers in other domains (inhibitory control), which is consistent with a foundational role of WM capacity for the dynamic process of skill formation (Cunha and Heckman 2007). Finally, we show that improvements in these domains can have causal, long-term effects on educational trajectories.

We believe that our approach has the advantages that (i) the children in the control group are participating in their normal school lessons, that is, we have a 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

[2] Several authors in the psychology and education science literature (Holmes, Gathercole, and Dunning 2009; St Clair-Thompson et al. 2010; Nutley and Söderqvist 2017) have also pointed out that, while direct effects of WM training on untrained WM tasks may happen in the short run, training-induced improvements in WM capacity need time to affect spillover outcomes.

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 be assessed. Unfortunately, these two studies experienced large attrition after a few months and/or did not have long-term follow-up measurements. In the light of our finding that many treatment effects become fully visible only after many months, this may have limited their ability to discover spillover effects.[3]

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 2013–14 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 consent from the parents of 580 (consent rate of 96.8%) for four waves of data collection (W1, W2, W3, W4). 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).[4] 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 below. About 49% of the children were male, and 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

[3] 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 3 months after the experiment.

[4] 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 on 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.

TABLE 1
Summary Statistics

Variable	Mean	SD	Minimum	Maximum	N
WM training	.488	.5	0	1	572
Male	.49	.5	0	1	572
Children's age (months):					
January 1, 2013	82.129	4.324	72.222	101.578	572
On test day W1	84.247	4.377	74.523	103.485	572
On test day W2	87.288	4.355	77.745	106.706	572
On test day W3	92.368	4.379	82.774	111.703	544
On test day W4	99.582	4.381	90.467	118.836	531
Migration background	.451	.498	0	1	568
Language problems	.247	.431	0	1	572
Monthly household net income (€):					
<750	.023	.149	0	1	441
750–1,500	.12	.326	0	1	441
1,500–2,500	.209	.407	0	1	441
2,500–5,000	.433	.496	0	1	441
>5,000	.215	.412	0	1	441
Mother's highest degree:					
University	.446	.498	0	1	444
Vocational	.423	.495	0	1	444
None	.131	.337	0	1	444
Secondary school type:					
Academic track	.692	.462	0	1	393
Mixed track	.204	.403	0	1	393
Nonacademic track	.104	.306	0	1	393

Note.—The table provides sociodemographic 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 knowing the migration background. 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 3 years after treatment.

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 groups; see app. 1.1 [appendix is available online]).

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 for 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 have 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. In addition, we did not inform parents about the treatment assignment of their children, and we also did not provide information that would have enabled them to infer the treatment assignment.[5]

We used a commercially available WM-training software providing training on different span tasks,[6] using an age-specific user interface and adaptive levels of difficulty. Eight out of 10 training tasks focus on visuospatial WM, while only two focus on verbal WM; that is, a much larger variety of WM tasks and more training time were allocated to visuospatial WM training. The teachers supervised children in each training session, and logins for the training software were user specific and valid only during the intervention period. Thus, the children had access to the training software only during their dedicated training sessions (see app. 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 covered in the first and second lessons of the day for first graders in primary school (mostly major subjects such as math and German language). This means that students in the treatment group missed 25 such 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, that is, whether computer-based WM training during school hours is beneficial for the children. In other words, when we compare the treatment- and control-group children on the various skill dimensions, we automatically take the forgone school lessons during WM training—that is, 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.[7]

[5] For further details on the information received by the parents, see app. 1.2.

[6] We used the WM-training software Cogmed. Cogmed and Cogmed Working Memory Training are trademarks, in the United States and/or other countries, of Cogmed Inc. (www.cogmed.com).

[7] Part of the literature on WM training emphasizes the importance of so-called active control groups. In our case, the control group is involved in the normal teaching lessons. It is

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).

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; see app. 1.3 for further details). Parents of both treatment- and control-group 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 this study. The entire sequence of tests was computer based, including auditory explanations (via headphones) 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 before 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, they received a small toy for participating in the evaluation waves. These rewards were given to all children from the control and treatment groups to avoid any motivational differences between them.

In each evaluation wave, the children completed three (nontrained) WM tasks. WM capacity was measured with a verbal simple span task, a

sometimes also argued that an active control group might perform nonadaptive WM training, i.e., the children are not exposed to increasingly challenging tasks when they have solved the less challenging ones. However, one disadvantage of nonadaptive 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 of whether WM training enables improvements relative to normal teaching lessons, our control group is involved in normal teaching lessons that typically involve increasingly challenging material.

verbal complex span task, and a visuospatial complex span task (for details, see app. 1.4). Importantly, both the verbal complex span task and the visuospatial complex span task clearly differed from the tasks used in the WM training. We included a verbal simple span task (but not a visuospatial simple span task) in the set of our WM evaluation tasks because the WM training places considerably less weight on verbal than on visuospatial WM. Direct effects may therefore be weaker for verbal WM. The verbal simple span task might allow us to capture these presumably weaker effects. The three WM tasks mentioned above not only enable us to study direct effects but also serve the purpose of examining the extent to which WM capacity mediates training-induced improvements in other important skills.

In each evaluation wave, the children also completed a set of tasks that enabled us to measure such spillover effects. 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, physics). In addition, we measured three other important skills that capture key aspects of EFs, such as fluid IQ (higher-level EFs), the ability to inhibit prepotent responses (inhibitory control), and the ability to sustain attention and display frustration tolerance (attentional stamina). We use Raven's Colored Progressive Matrices test (Bulheller and Häcker 2010) as a measure for fluid IQ. The ability to inhibit prepotent responses (inhibitory control) was measured with the go/no-go task (Gawrilow and Gollwitzer 2008), and attentional stamina was measured using the bp (letter discrimination) task (Esser, Wyschkon, and Ballaschk 2008). For a detailed description of all these tasks, see appendix 1.4.

2. Teacher Ratings

In each data collection wave (W1–W4), teachers filled out a questionnaire containing items on children's and teachers' characteristics and behaviors, and (for treated teachers) expectations about the intervention. 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 app. 1.4).

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 5, that is, 3 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 nonacademic track (*Realschule Plus*). In this particular federal state in Germany, 86% of the children in the academic track earn a degree that qualifies them for general university enrollment (*Abitur*), whereas only 25% of children in mixed-track schools achieve this (Rhineland-Palatinate Statistics Office 2018). Within the nonacademic track, students cannot earn a degree that qualifies them for general university enrollment. For children in the nonacademic track, the probability of switching tracks is small (<5% per year; Bellenberg 2012). Moreover, since the 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 lifetime earnings.

D. Conjectures about the Treatment Effect on Outcome Measures

In addition to direct effects on WM capacity, WM training may have positive spillover effects on our educational outcome measures, 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. However, in our context it is 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 on teaching arithmetic. Because the treatment subjects missed a considerable number of math lessons and because our WM training was focused on visuospatial WM (see above), it seems more likely that we would 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 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 Söderqvist 2017) that WM capacity does not predict word identification but 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 spillover effects generally may need time to emerge—for why WM-training effects in our reading task may emerge only over time.

Turning to more general cognitive skills, WM capacity has also been shown to be correlated with fluid intelligence as measured, for example, by Raven's matrices task—a task that requires visuospatial WM but is nevertheless different from pure WM tasks because it requires (i) reasoning in novel situations without prior knowledge and (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).[8] 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-Lervåg, Redick, and Hulme 2016).

WM is one of three core components of executive functioning—with inhibitory control and cognitive flexibility being the other two (Diamond 2013). The literature on executive functioning hypothesizes that WM and inhibitory control "generally need one another and co-occur" and that WM "supports inhibitory control" (Diamond 2013, 143). This is also consistent with the view that WM capacity is crucial for the ability to actively maintain task-relevant and suppress/inhibit task-irrelevant information (Engle 2002). WM capacity might thus 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). On the basis of this account, WM training may thus generate spillover effects on inhibitory control. In the context of the go/no-go task, this means that children who undergo WM training should be better able to avoid commission errors, 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. We also measure children's attentional stamina with a letter discrimination task, the so-called bp task. To our knowledge, it is an open question whether WM training improves this aspect of EFs.[9]

[8] Note that Raven's matrices task does not measure general IQ but is a nonverbal test that is regarded as a measure of fluid intelligence based on visuospatial capabilities.

[9] Our WM training may also be viewed from the perspective of prominent interventions that boosted executive functioning (see, e.g., Walker et al. 2022; García, Heckman, and Ronda 2023) and led to long-lasting spillover effects on a wide range of skills.

Finally, in case we find that WM training has spillover effects on academic performance or other important skills, it might be possible that WM training also positively affects secondary school track choice because that choice is presumably influenced by children's academic skills, their fluid IQ, and their self-regulatory 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.[10] All outcome scores are standardized within each evaluation wave to mean 0 and SD 1. We control for the pretraining 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; that is, 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 step-down procedure to control for multiple-hypothesis testing (Romano and Wolf 2005, 2016)—a technique that is increasingly used for large-scale intervention studies (see, e.g., Cunha, Heckman, and Schennach 2010, Campbell et al. 2014, and 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 (bias-reduced linearization) correction method (Bell and McCaffrey 2002).

A. *Sample Balance*

To examine whether randomization led to a balanced sample across treatment and control groups in terms of socioeconomic characteristics, we regress various sociodemographic characteristics (gender, age, and migration background, as well as parental income and education) measured before the treatment (W1) on the treatment indicators and school fixed effects (see table S1; tables S1–S23 are available online). The results show that the treatment coefficient in all regressions is close to zero

[10] The vector of control variables consists of school fixed effects, gender, age, age on test days, baseline value of the outcome, and indicators for other treatments (unrelated to the WM training) that were conducted in the same sample. For further details on estimation, see app. 1.5).

and insignificant, indicating that there were no significant imbalances between treatment and control groups with respect to these variables.

As a further sample balance check, we regressed standardized outcome test scores at baseline (i.e., test scores measured before the treatment, in W1) on the treatment indicators, school fixed effects, and the same control variables that are included in the main estimations of the treatment effect. Table S2 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 groups. 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 BRL clustering method (which accounts for small numbers of clusters). This then yields a p-value of .332 for the verbal complex span outcome, suggesting no significant difference between the treatment and control groups 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.

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 and controls (see above). The estimated direct effects of WM training on WM capacity are presented in figure 1 and table S3. We find significantly positive treatment effects for the visuospatial complex span task in all three posttreatment waves, with an effect size (d) of 0.40–0.46 SD ($p = .00004$–.006). We also find a significantly positive training effect on performance in the verbal simple span task of $d = 0.38$ SD ($p = .000008$) in W3 and $d = 0.30$ SD ($p = .015$) in W4. We do not find any significant treatment effect for performance in the verbal complex span task. The stronger effect of training on visuospatial WM compared to verbal WM is plausible, as the training focused primarily on visuospatial WM (see sec. II.B).

Spillover effects of WM training on educational outcomes—arithmetic, geometry, and reading—and Raven's fluid-IQ measure are reported in figure 2 and table S4. While there is no treatment effect on arithmetic in all three posttraining waves, we find an effect on geometry skills that is increasing over time. The effect size $d = 0.17$ SD in W2 is not yet significantly different from zero ($p = .108$), but the effect size increases in W3 and W4 to $d = 0.24$ and $d = 0.38$ SD, respectively, with significance levels of

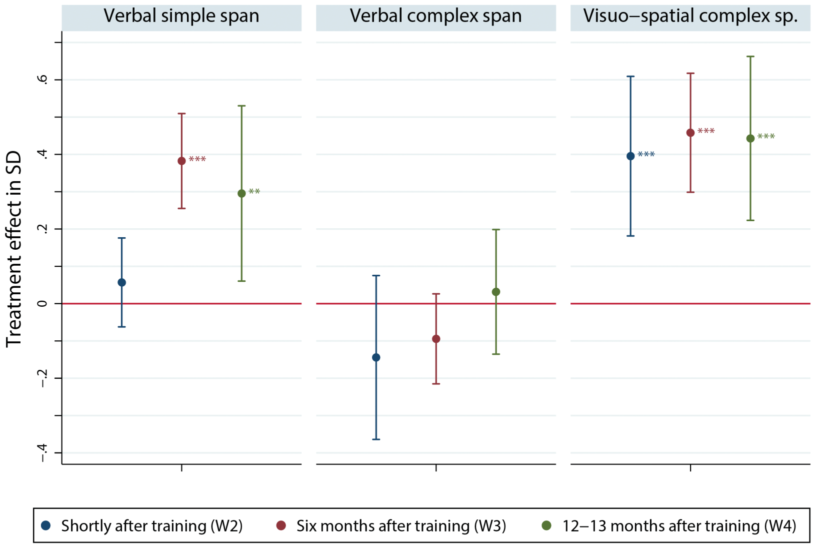

FIG. 1.—Direct effect of training on WM capacity. The dots show the point estimates (as fractions of an SD) of how WM training changes the performance in the three WM tasks (indicated in the subfigure title) relative to the control group. The bars indicate the 95% confidence intervals. All estimates are based on least squares models controlling for school fixed effects, pretreatment outcome scores, and further controls (see app. 1.5 for details). The econometric estimates are shown in table S3. The confidence intervals and the associated significance statements are based on the clustering of standard errors at the classroom level. $^{**}p < .05$, $^{***}p < .01$.

$p = .021$ in W3 and $p = .001$ in W4. Thus, it seems that WM training had a positive and increasing spillover effect relative to the normal school curriculum on geometry skills but not on arithmetic skills. The significant and relatively strong impact on geometry skills is also consistent with the fact that training focused primarily on improving visuospatial WM capacity. The spillover effects on reading are generally lower than those on 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 $d = 0.23$ SD at $p = .037$ in W4. This rising spillover effect on reading is consistent with the view (Nutley and Söderqvist 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 also find a significant spillover effect on Raven's Colored Matrices task 6 months ($d = 0.24$ SD, $p = .004$) and 12–13 months ($d = 0.24$ SD, $p = .002$) after the training. We emphasize that this finding does not mean that WM training increased all dimensions of fluid intelligence,

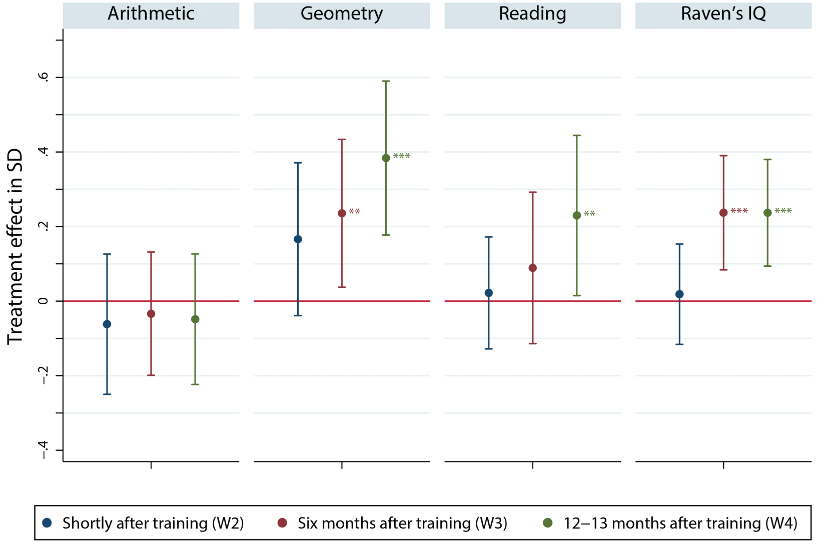

Fig. 2.—Spillover effects on arithmetic, geometry, reading, and Raven's IQ. The dots show the point estimates (as fractions of an SD) of how WM training changes performance in arithmetic, geometry, reading, and Raven's fluid-IQ measure relative to the control group. The bars indicate the 95% confidence intervals. All estimates are based on least squares models controlling for school fixed effects, pretreatment outcome scores, and further controls (see app. 1.5 for details). The econometric estimates are shown in table S4. The confidence intervals and the associated significance statements are based on the clustering of standard errors at the classroom level. $^{**}p < .05$, $^{***}p < .01$.

as some research indicates that only 64% of the variance in performance in the Raven task is attributable to general fluid intelligence (Jensen 1998). However, the Raven task measures important dimensions of fluid intelligence that require WM capacity (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 fluid-IQ measure are driven by a decline in the performance of the control group. As a result of cognitive maturation over the course of 1 year, both the treatment and control groups increased their performance over time. Accordingly, 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 (fig. 3; table S5). We find positive spillover effects of WM training on children's inhibitory control measured in the go/no-go task. We measure inhibitory control by multiplying children's standardized number of commission errors by -1; that is, a reduction in commission errors shows up as a numerical increase in this performance measure. Figure 3 indicates a highly significant reduction in commission errors in the treatment relative to the control group in W4 ($d = 0.33$ SD,

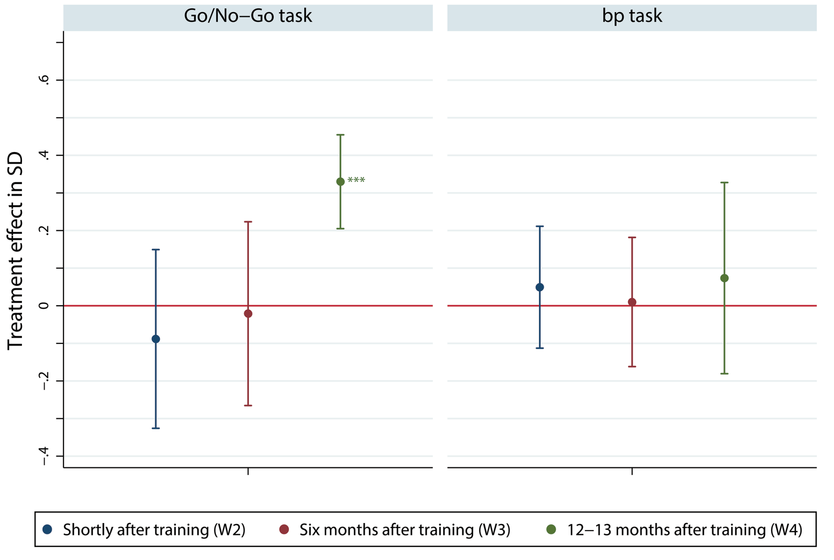

FIG. 3.—Spillover effects in the go/no-go task and the bp task. The dots show the point estimates (as fractions of an SD) of how WM training changes the performance in the tasks relative to the control group. The bars indicate the 95% confidence intervals. All estimates are based on least squares models controlling for school fixed effects, pretreatment outcome scores, and further controls (see app. 1.5 for details). The econometric estimates are shown in table S5. The confidence intervals and the associated significance statements are based on the clustering of standard errors at the classroom level. $^{***}p < .01$.

$p < .0001$).[11] Interestingly, while we observe no treatment effect on commission errors in W2 and W3, we observe a weakly significant treatment effect on performance in terms of a reduction in response times in W2 ($d = 0.23$ SD, $p = .053$) and W3 ($d = 0.37$ SD, $p = .094$). 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.[12]

Overall, these data patterns suggest that, similar to the cases of geometry, reading, and Raven's fluid-IQ measure, spillover effects on inhibitory

[11] We also analyze 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—and find a significant performance effect in W4 (W2: $d = 0.118$ SD, $p = .410$; W3: $d = 0.071$ SD, $p = .619$; W4: $d = 0.475$ SD, $p < .0001$). If we analyze omission errors (i.e., failing to push the button in go trials, which is often interpreted as a measure for "attention") separately, we also find similar positive treatment effects as for inhibitory control, with the strongest and most significant improvements in W4 (W2: $d = 0.282$ SD, $p = .109$; W3: $d = 0.133$ SD, $p = 0.357$; W4: $d = 0.416$ SD, $p = .001$).

[12] Similarly, when analyzing teacher-reported overall self-regulation as a measure of everyday self-regulatory behavior in the classroom, we also find significant positive treatment effects (see apps. 1.4 and 1.5 and table S19).

control emerge over time. This effect supports the theoretical conjecture that WM "supports inhibitory control" (Diamond 2013, 143). Note also that this spillover effect is due to a differentially larger increase in the performance of the treatment group relative to the control group in terms of 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.

C. *Treatment Effect on Choice of Secondary School Track*

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

Controlling for the same set of variables as for the other treatment effects, we find that children in the treatment group are roughly 16 percentage points more likely to choose the advanced track of secondary school, relative to children in the control group (table 2, col. 1). If we estimate the treatment effect with a probit model instead of a linear probability model (table 2, col. 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 advanced track of secondary school. If we take the full range of secondary school choices (advanced track, mixed track, nonacademic track) into account, we again find a sizeable positive treatment effect on enrollment in the advanced track (cols. 3 and 4). Column 4 of table 2 also indicates that the increase in advanced-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 nonacademic-track enrollment.

As we measure the secondary school track enrollment 3 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 groups. In the treatment group, we still can collect data of 68.1% of the sample in W1, and in the control group we have data of 69.3% of the sample in W1 (see app. 1.6 for further robustness checks on attrition).

We also address systematic attrition by estimating inverse-probability weighting models. To apply these models, we compared the sample characteristics in W1 with the sample characteristics at the time of secondary

TABLE 2
Treatment Effect of WM Training on Secondary School Choice at Age 10 ($N = 393$)

Secondary School Choice	OLS (1)	Probit (2)	OLS Categorical Variable (3)	Ordered Probit (4)	Inverse-Probability Weighting (5)
Academic track	.157***	.148***	.221***	.136***	.170***
	(.050)	(.045)	(.078)	(.046)	(.050)
Mixed track				−.067***	
				(.025)	
Nonacademic track				−.069***	
				(.023)	

Note.—Column 1 reports the effect of the treatment on the probability of being enrolled in an academic-track secondary school, based on an ordinary least squares (OLS) model. When we cluster the standard errors using BRL, the standard error in col. 1 becomes 0.070 (which corresponds to a p-value of .026). Column 2 reports the marginal treatment effect of the probit estimate on the same dependent variable as in col. 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 nonacademic-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 advanced-track school (*Gymnasium*). Column 4 reports the marginal treatment effects of the ordered probit estimates on the same dependent variable as in col. 3. Column 5 reports an estimation similar to that in col. 1 but accounts for attrition by applying inverse-probability weighting. The weights are calculated for groups defined on the basis of migration background, high/low academic performance (math and reading performance), and high/low cognitive performance (WM capacity and Raven's fluid-IQ measure). All models include school fixed effects and further controls (see app. 1.5 for further details, including our calculation of the inverse-probability weights). Standard errors, in parentheses, are clustered at the classroom level.
*** $p < .01$.

school choice. This comparison shows that at the time of secondary school choice there are (i) fewer children with a migration background, (ii) more children with higher academic performance (i.e., geometry, arithmetic, and reading), and (iii) more children with higher cognitive skills (i.e., WM capacity and Raven's fluid-IQ measure). Therefore, we calculated the inverse-probability weights for groups defined on the basis of three binary variables: (i) migration background, (ii) high/low academic performance in geometry, arithmetic, and reading, and (iii) high/low cognitive skills as measured by WM capacity and Raven's fluid-IQ measure. The result of this model (shown in col. 5) indicates that the WM training increases advanced-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% chose the advanced track; for those whose mother does not have a university degree, the number is 54%, that is, a difference of 32 percentage points. This difference declines to 27 percentage points when children's baseline measure of Raven's fluid IQ is controlled for. Thus, the 14–17 percentage point increase in advanced-track enrollment is substantial when compared with this socioeconomic gap.

D. Heterogeneous 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 tables S6–S8). The results show that children with low baseline WM capacity perform substantially worse in all spillover 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 treatment effect is not systematically different for children with low WM capacity. Importantly, however, the treatment 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.

E. Robustness Checks

We perform a series of robustness checks, including checks for attrition, the potential role of computer use, Hawthorne or demand-type effects, and multiple-hypothesis testing corrections. For the multiple-hypothesis testing, we grouped our outcomes into four families, following the above conjectures for treatment effects: (1) WM outcomes (verbal simple span, verbal complex span, visuospatial complex span), (2) spillover effects on educational outcomes (arithmetic, geometry, reading), (3) spillover effects on general cognitive skills (Raven's IQ), and (4) spillover effects on general noncognitive skills (go/no-go task, bp task). Note that each family includes three measurements for each outcome (at W2, W3, and W4). Overall, these robustness checks confirm our findings, except for the treatment effect on reading in W4, which turns insignificant if we correct for multiple-hypothesis testing (see table S9).[13] All details on robustness can be found in appendix 1.6.

[13] We also provide further multiple-hypothesis testing analyses in table S10, using an even more conservative grouping into only two families (direct effects and spillover

IV. Mechanisms

In our view, the documented treatment effects on WM capacity and on spillover outcomes have a plausible interpretation. For example, it is plausible that WM training has an immediate effect on visuospatial WM capacity (i.e., the aspect of WM that received the most emphasis during the training), while spillover effects need more time to evolve—which is what we observe in our data. Likewise, the finding that WM training does not increase arithmetic but does increase geometry skills may be due to the fact that the training emphasized visuospatial WM, which may well play a larger role in geometry than in arithmetic. Similarly, visuospatial WM capacity is likely to be a basic prerequisite to deploy the problem-solving skill that is required to solve Raven's fluid-IQ task.

To provide a quantitative assessment of the extent to which WM capacity might be a mediating mechanism for the observed spillover effects, we performed a mediation analysis by using the method applied in Heckman, Pinto, and Savelyev (2013), and similarly in, for example, Carlana, La Ferrara and Pinotti (2018) and Kosse et al. (2020). The formal details of this method are described in appendix 1.5. Intuitively, the method provides us with the share of the total treatment effect of the training on each spillover outcome that can be explained by the training-induced changes in WM capacity.

The results of our mediation analysis are presented in figure 4. The figure shows that for geometry, reading, and Raven's fluid-IQ measure, a large part of the total treatment effect—roughly between 50% and 66%—is mediated by WM capacity. Interestingly, the mediation effect of WM capacity is much lower for our measure of inhibitory control (performance in the go/no-go task). Perhaps this lower mediation effect of WM capacity is one reason why the training effect on the ability to inhibit prepotent impulses took more time to develop.

Overall, this analysis suggests that training-induced changes in WM capacity appear to explain substantial parts of the treatment effect on spillover outcomes. In view of the previous literature on WM training (e.g., Sala et al. 2019; Sala and Gobet 2020), we were, however, surprised by the magnitude of the effects on spillover outcomes. Therefore, we point out specificities of our study that are likely to be relevant in this context. First, we delivered the WM training in a school context as part of the regular curriculum, which ensures high external validity. Moreover, the integration of the treatment into regular classroom teaching may have facilitated the

effects). Again, three measurements are included for each outcome in a family (at W2, W3, and W4). While we believe that the grouping of families described in this subsection is the most reasonable, the choice of families of outcomes is always somewhat discretionary. With the very conservative grouping of outcomes into only two families, results remain similar to those in table S9, but the treatment effects on Raven's IQ are no longer significant at conventional levels (W3: $p = .136$, W4: $p = .114$).

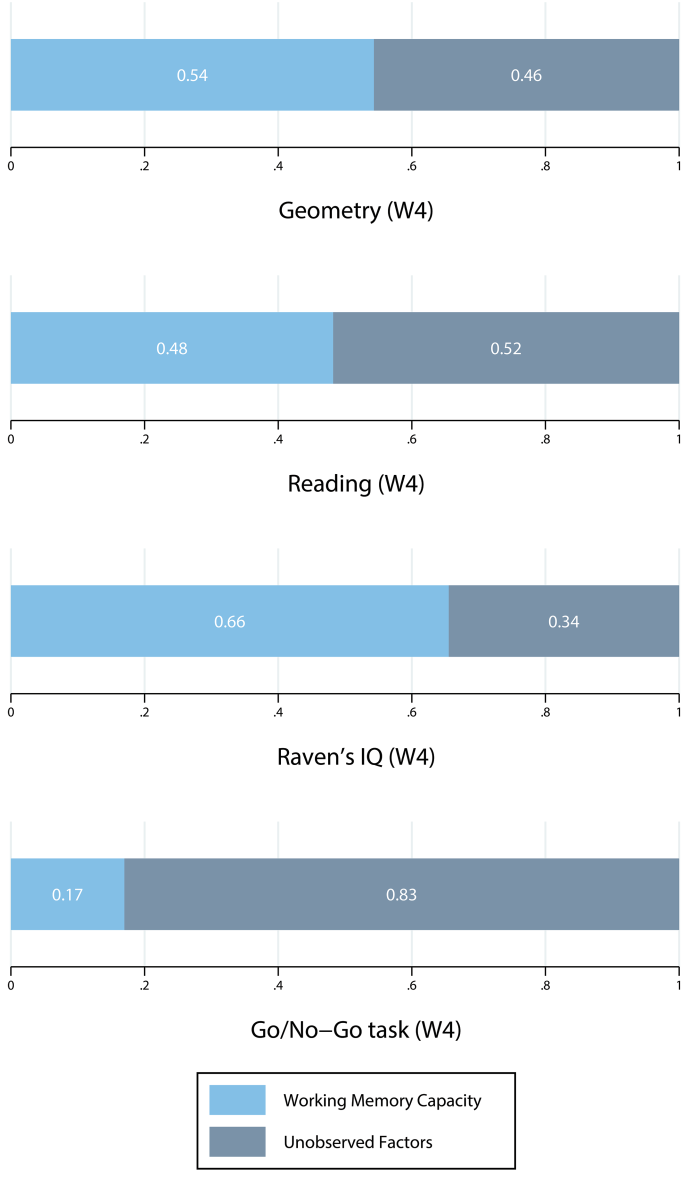

FIG. 4.—Relative importance of WM capacity for treatment effects on spillover outcomes. This figure displays the estimated decomposition of the total treatment effect on those spillover outcomes that are significantly improved by the WM training in W4 (12–13 months after treatment). For each outcome, we estimate the effect of the treatment that is mediated by WM capacity (see app. 1.5 for details). Light blue bars show the percentage of the treatment effect that is mediated by training-induced increases in WM capacity.

spillover effects on other school-related skills. The context of regular classroom teaching is also likely to minimize placebo or Hawthorne effects. Of course, we cannot rule out the possibility that placebo effects may have played a role, for example, because the children in the treatment group received extra attention (e.g., because they used computers in class or because of the presence of a research assistant during the training; see app. 1.2). Moreover, by design minor differences between treatment and control groups inevitably remain that could potentially affect abilities other than WM capacity (e.g., narrow task learning due to familiarity with WM tasks).[14] However, the facts that (i) we carefully developed outcome measures that are different from the training tasks (even with respect to input devices, i.e., touchscreens vs. external mice; see sec. II.C), that (ii) we see treatment effects on very specific spillover outcomes that require visuospatial WM (and no effect on other important educational outcomes), and that (iii) the pattern of effects increases over time suggest that placebo effects or remaining minor group differences are unlikely to have played a substantial role.[15]

Second, as mentioned above, our study may be better capable of detecting spillover effects because we also measure the relevant outcomes 6 and 12–13 months after the treatment, while most other studies stop collecting spillover outcomes after a few months and thus cannot identify effects that might take a longer time to evolve. Third, because we treated complete classes (class-wise randomization), in addition to effects on individual-level skills, the treatment possibly led to various sorts of positive peer-group and classroom effects that, in turn, could have affected teachers' behavior and attitudes. In our setting, such beneficial peer-group effects seem plausible, given that the children usually stay together in the same class and with the same teacher for 4 years in primary school. Evidently, these peer-group effects constitute an important factor for the persistence of treatment effects of interventions at young ages (see Bailey et al. 2017).

V. Summary

On the basis of a randomized controlled trial with 572 first graders in primary schools, we found that a 5-week, one-lesson-per-school-day, adaptive

[14] For example, in a computer-based WM training, treated children will automatically become more familiar with WM tasks. Thus, they may perform better in subsequent WM tasks merely because they are more familiar with the type of tasks and not because they have higher WM capacity. Similarly, they have more screen time than children in the control group, which could potentially improve skills such as perceptual speed.

[15] Note also that our intervention was part of a larger educational study, involving other treatments (see Schunk et al. 2022). However, we control for the other treatments in all our estimations, and we conduct various robustness checks, including correction for multiple-hypothesis testing and the small number of clusters, to minimize the likelihood of false positives or spurious findings (for details, see apps. 1.5 and 1.6).

WM training during class not only improves children's WM capacity but also has spillover effects on their geometry skills, Raven's fluid-IQ measure, and ability to inhibit prepotent impulses. We observe an increasing pattern of treatment effects on these spillover 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 considerable parts of the spillover effects. When assessing the reported effect sizes for the spillover effects, it is interesting to compare them with effect sizes observed in other (more intensive) intervention studies—such as the Perry Preschool Project, the Jamaican supplementation and stimulation study, and others—producing improvements in EFs even in the very long run of 0.25 SD to well above 0.5 SD (Riggs et al. 2006; Raver et al. 2011; Heckman, Pinto, and Savelyev 2013; Gertler et al. 2014; Walker et al. 2022; García, Heckman, and Ronda 2023). Finally, we document that the WM training can have an impact on one of the most consequential school career decisions in the German school system: whether to enroll the child in the advanced track of secondary school (*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 significantly higher salaries. The increasing pattern of effects on spillover outcomes combined with the effect on long-run educational choices is consistent with the idea of self-productivity in the process of skill formation (Cunha and Heckman 2007). Taken together, our findings thus provide novel evidence consistent with the dynamic process of skill formation, and they suggest that our treatment generated substantial benefits for the children.

Data Availability

The data for this publication have been collected in a project that has compiled a large set (and combination) of children's abilities, preferences, and family (sociodemographic) characteristics (see apps. 1.3 and 1.4) and thus represents highly sensitive data. This dataset cannot be made available for data protection reasons. In addition, parental consent for data usage covers only strictly scientific purposes. The restriction to scientific purposes was also necessary to comply with data protection requirements, and use of the data for strictly scientific purposes cannot be guaranteed if the dataset is made (publicly) available. Not all the data collected in this project are analyzed for this publication; see appendix 1.4 for details. Researchers interested in replicating our findings can get access to the dataset after filling out a research agreement with our universities. The code replicating the tables and figures in this article can be found in Berger et al. (2024) in the Harvard Dataverse, https://doi.org/10.7910/DVN/PQNQU0. We confirm that in the paper and the appendix, we have reported all measures, conditions, and data exclusions and how we determined our sample sizes.

References

Ackerman, Phillip L., Margaret E. Beier, and Mary O. Boyle. 2005. "Working Memory and Intelligence: The Same or Different Constructs?" *Psychological Bull.* 131 (1): 30–60.

Aksayli, N. Deniz, Giovanni Sala, and Fernand Gobet. 2019. "The Cognitive and Academic Benefits of Cogmed: A Meta-Analysis." *Educ. Res. Rev.* 27:229–43.

Alan, Sule, and Seda Ertac. 2018. "Fostering Patience in the Classroom: Results from Randomized Educational Intervention." *J.P.E.* 126 (5): 1865–911.

Alloway, Tracy Packiam, and Ross G. Alloway. 2010. "Investigating the Predictive Roles of Working Memory and IQ in Academic Attainment." *J. Experimental Child Psychology* 106 (1): 20–29.

Almond, Douglas, Janet Currie, and Valentina Duque. 2018. "Childhood Circumstances and Adult Outcomes: Act II." *J. Econ. Literature* 56 (4): 1360–446.

Au, Jacky, Ellen Sheehan, Nancy Tsai, Greg J. Duncan, Martin Buschkuehl, and Susanne M. Jaeggi. 2015. "Improving Fluid Intelligence with Training on Working Memory: A Meta-Analysis." *Psychonomic Bull. and Rev.* 22 (2): 366–77.

Baddeley, Alan D. 1999. *Essentials of Human Memory.* Hove, UK: Psychology.

Bailey, Drew, Greg J. Duncan, Candice L. Odgers, and Winnie Yu. 2017. "Persistence and Fadeout in the Impacts of Child and Adolescent Interventions." *J. Res. Educ. Effectiveness* 10 (1): 7–39.

Bell, Robert M., and Daniel F. McCaffrey. 2002. "Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples." *Survey Methodology* 28 (2): 169–81.

Bellenberg, Gabriele. 2012. *Schulformwechsel in Deutschland.* Gütersloh, Germany: Bertelsmann Stiftung.

Berger, Eva M., Ernst Fehr, Henning Hermes, Daniel Schunk, and Kirsten Winkel. 2024. "Replication Data for: 'The Impact of Working Memory Training on Children's Cognitive and Noncognitive Skills.'" Harvard Dataverse, https://doi.org/10.7910/DVN/PQNQU0.

Borghans, Lex, Angela Lee Duckworth, James J. Heckman, and Bas ter Weel. 2008. "The Economics and Psychology of Personality Traits." *J. Human Resources* 43 (4): 972–1059.

Bulheller, Stephan, and Hartmut O. Häcker. 2010. *Coloured Progressive Matrices (CPM): Deutsche Bearbeitung und Normierung nach J. C. Raven, J. Raven und J. H. Court.* Frankfurt: Pearson Assessment.

Campbell, Frances, Gabriella Conti, James J. Heckman, Seong Hyeok Moon, Rodrigo Pinto, Elizabeth Pungello, and Yi Pan. 2014. "Early Childhood Investments Substantially Boost Adult Health." *Science* 343 (6178): 1478–85.

Cappelen, Alexander, John List, Anya Samek, and Bertil Tungodden. 2020. "The Effect of Early-Childhood Education on Social Preferences." *J.P.E.* 128 (7): 2739–58.

Carlana, Michela, Eliana La Ferrara, and Paolo Pinotti. 2018. "Goals and Gaps: Educational Careers of Immigrant Children." HKS Faculty Research Working Paper RWP18-036 (August), Harvard Kennedy School, Cambridge, MA.

Carpenter, Patricia A., Marcel A. Just, and Peter Shell. 1990. "What One Intelligence Test Measures: A Theoretical Account of the Processing in the Raven Progressive Matrices Test." *Psychological Rev.* 97 (3): 404–31.

Constantinidis, Christos, and Torkel Klingberg. 2016. "The Neuroscience of Working Memory Capacity and Training." *Nature Rev. Neuroscience* 17 (7): 438–49.

Conti, Gabriella, and James J. Heckman. 2010. "Understanding the Early Origins of the Education-Health Gradient: A Framework That Can Also Be Applied to Analyze Gene-Environment Interactions." *Perspectives Psychological Sci.* 5 (5): 585–605.

Cunha, Flavio, and James Heckman. 2007. "The Technology of Skill Formation." *A.E.R.* 97 (2): 31–47.

———. 2009. "The Economics and Psychology of Inequality and Human Development." *J. European Econ. Assoc.* 7 (2–3): 320–64.

Cunha, Flavio, James J. Heckman, Lance Lochner, and Dimitriy V. Masterov. 2006. "Interpreting the Evidence on Life Cycle Skill Formation." In *Handbook of the Economics of Education*, vol. 1, edited by Eric A. Hanushek and Finis Welch, 698–812. Amsterdam: North-Holland.

Cunha, Flavio, James J. Heckman, and Susanne M. Schennach. 2010. "Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Econometrica* 78 (3): 883–931.

Diamond, Adele. 2013. "Executive Functions." *Ann. Rev. Psychology* 64:135–68.

Diamond, Adele, and Daphe S. Ling. 2020. "Review of the Evidence on, and Fundamental Questions about, Efforts to Improve Executive Functions, Including Working Memory." In *Cognitive and Working Memory Training: Perspectives from Psychology, Neuroscience, and Human Development*, edited by Jared M. Novick, Michael F. Bunting, Michael R. Dougherty and Randall W. Engle, 143–431. New York: Oxford Univ. Press.

Duckworth, Angela L. 2011. "The Significance of Self-Control." *Proc. Nat. Acad. Sci. USA* 108 (7): 2639-40.

Duckworth, Angela L., and Stephanie M. Carlson. 2013. "Self-Regulation and School Success." In *Self-Regulation and Autonomy: Social and Developmental Dimensions of Human Conduct*, edited by Bryan W. Sokol, Frederick M. E. Grouzet, and Ulrich Müller, 208–30. Cambridge: Cambridge Univ. Press.

Duckworth, Angela L., Christopher Peterson, Michael D. Matthews, and Dennis R. Kelly. 2007. "Grit: Perseverance and Passion for Long-Term Goals." *J. Personality and Soc. Psychology* 92 (6): 1087–101.

Duckworth, Angela L., David Weir, Eli Tsukayama, and David Kwok. 2012. "Who Does Well in Life? Conscientious Adults Excel in Both Objective and Subjective Success." *Frontiers Psychology* 3. https://doi.org/10.3389/fpsyg.2012.00356.

Dustmann, Christian. 2004. "Parental Background, Secondary School Track Choice, and Wages." *Oxford Econ. Papers*, n.s., 56 (2): 209–30.

Dweck, Carol S. 2006. *Mindset: The New Psychology of Success.* New York: Random House.

Engel de Abreu, Pascale M. J., Andrew R. A. Conway, and Susan E. Gathercole. 2010. "Working Memory and Fluid Intelligence in Young Children." *Intelligence* 38 (6): 552–61.

Engle, Randall W. 2002. "Working Memory Capacity as Executive Attention." *Current Directions Psychological Sci.* 11 (1): 19–23.

Esser, Günter, Anne Wyschkon, and Katja Ballaschk. 2008. *Basisdiagnostik umschriebener Entwicklungsstörungen im Grundschulalter (BUEGA).* Göttingen: Hogrefe.

Frison, Lars, and Stuart J. Pocock. 1992. "Repeated Measures in Clinical-Trials: Analysis Using Mean Summary Statistics and Its Implications for Design." *Statis. Medicine* 11 (13): 1685–704.

García, Jorge Luis, and James J. Heckman. 2023. "Parenting Promotes Social Mobility within and across Generations." *Ann. Rev. Econ.* 15:349–88.

García, Jorge Luis, James J. Heckman, and Victor Ronda. 2023. "The Lasting Effects of Early-Childhood Education on Promoting the Skills and Social Mobility of Disadvantaged African Americans and Their Children." *J.P.E.* 131 (6): 1477–506.

Gaspar, John M., Gregory J. Christie, David J. Prime, Pierre Jolicoeur, and John J. McDonald. 2016. "Inability to Suppress Salient Distractors Predicts Low Visual Working Memory Capacity." *Proc. Nat. Acad. Sci. USA* 113 (13): 3693–98.

Gathercole, Susan E., Susan J. Pickering, Camilla Knight, and Zoe Stegmann. 2004. "Working Memory Skills and Educational Attainment: Evidence from National Curriculum Assessments at 7 and 14 Years of Age." *Appl. Cognitive Psychology* 18 (1): 1–16.

Gawrilow, Caterina, and Peter M. Gollwitzer. 2008. "Implementation Intentions Facilitate Response Inhibition in Children with ADHD." *Cognitive Therapy and Res.* 32 (2): 261–80.

Gertler, Paul, James Heckman, Rodrigo Pinto, Arianna Zanolini, Christel Vermeersch, Susan Walker, Susan M. Chang, and Sally Grantham-McGregor. 2014. "Labor Market Returns to an Early Childhood Stimulation Intervention in Jamaica." *Science* 344 (6187): 998–1001.

Gobet, Fernand, and Giovanni Sala. 2023. "Cognitive Training: A Field in Search of a Phenomenon." *Perspectives Psychological Sci.* 18 (1): 125–41.

Green, C. Shawn, Daphne Bavelier, Arthur F. Kramer, Sophia Vinogradov, Ulrich Ansorge, Karlene K. Ball, Ulrike Bingel, et al. 2019. "Improving Methodological Standards in Behavioral Interventions for Cognitive Enhancement." *J. Cognitive Enhancement* 3 (1): 2–29.

Heckman, James J. 2006. "Skill Formation and the Economics of Investing in Disadvantaged Children." *Science* 312 (5782): 1900–1902.

Heckman, James, Rodrigo Pinto, and Peter Savelyev. 2013. "Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes." *A.E.R.* 103 (6): 2052–86.

Heckman, James J., Jora Stixrud, and Sergio Urzua. 2006. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *J. Labor Econ.* 24 (3): 411–82.

Hofmann, Wilhelm, Tobias Gschwendner, Malte Friese, Reinout W. Wiers, and Manfred Schmitt. 2008. "Working Memory Capacity and Self-Regulatory Behavior: Toward an Individual Differences Perspective on Behavior Determination by Automatic versus Controlled Processes." *J. Personality and Soc. Psychology* 95 (4): 962–77.

Holmes, Joni, Susan E. Gathercole, and Darren L. Dunning. 2009. "Adaptive Training Leads to Sustained Enhancement of Poor Working Memory in Children." *Developmental Sci.* 12 (4): F9–F15.

Jensen, Arthur R. 1998. *The G Factor: The Science of Metal Ability. Human Evolution, Behavior and Intelligence.* Westport, CT: Praeger.

Karbach, Julia, and Paul Verhaeghen. 2014. "Making Working Memory Work: A Meta-Analysis of Executive-Control and Working Memory Training in Older Adults." *Psychological Sci.* 25 (11): 2027–37.

Katz, Benjamin, and Priti Shah. 2016. "The Jury Is Still Out on Working Memory Training." *JAMA Pediatrics* 170 (9): 907–8.

Kibby, Michelle Y., Sylvia E. Lee, and Sarah M. Dyer. 2014. "Reading Performance Is Predicted by More Than Phonological Processing." *Frontiers Psychology* 5 (2014). https://doi.org/10.3389/fpsyg.2014.00960.

Klingberg, Torkel. 2016. "Neural Basis of Cognitive Training and Development." *Current Opinion Behavioral Sci.* 10:97–101.
Klingberg, Torkel, Elisabeth Fernell, Pernille J. Olesen, Mats Johnson, Per Gustafsson, Kerstin Dahlström, Christopher G. Gillberg, Hans Forssberg, and Helena Westerberg. 2005. "Computerized Training of Working Memory in Children with ADHD—A Randomized, Controlled Trial." *J. American Acad. Child and Adolescent Psychiatry* 44 (2): 177–86.
Kosse, Fabian, Thomas Deckers, Pia Pinger, Hannah Schildberg-Horisch, and Armin Falk. 2020. "The Formation of Prosociality: Causal Evidence on the Role of Social Environment." *J.P.E.* 128 (2): 434–67.
Kyllonen, Patrick C., and Raymond E. Christal. 1990. "Reasoning Ability Is (Little More Than) Working Memory Capacity?!" *Intelligence* 14 (4): 389–433.
Martinussen, Rhonda, Jill Hayden, Sheilah Hogg-Johnson, and Rosemary Tannock. 2005. "A Meta-Analysis of Working Memory Impairments in Children with Attention-Deficit/Hyperactivity Disorder." *J. American Acad. Child and Adolescent Psychiatry* 44 (4): 377–84.
McKenzie, David. 2012. "Beyond Baseline and Follow-Up: The Case for More T in Experiments." *J. Development Econ.* 99 (2): 210–21.
Melby-Lervåg, Monica, Thomas S. Redick, and Charles Hulme. 2016. "Working Memory Training Does Not Improve Performance on Measures of Intelligence or Other Measures of 'Far Transfer': Evidence from a Meta-Analytic Review." *Perspectives Psychological Sci.* 11 (4): 512–34.
Moffitt, Terrie E., Louise Arseneault, Daniel Belsky, Nigel Dickson, Robert J. Hancox, HonaLee Harrington, Renate Houts, et al. 2011. "A Gradient of Childhood Self-Control Predicts Health, Wealth, and Public Safety." *Proc. Nat. Acad. Sci. USA* 108 (7): 2693–98.
Nutley, Sissela Bergman, and Stina Söderqvist. 2017. "How Is Working Memory Training Likely to Influence Academic Performance? Current Evidence and Methodological Considerations." *Frontiers Psychology* 8:69. https://doi.org/10.3389/fpsyg.2017.00069.
Oberauer, Klaus, Ralf Schulze, Oliver Wilhelm, and Heinz-Martin Süß. 2005. "Working Memory and Intelligence—Their Correlation and Their Relation: Comment on Ackerman, Beier, and Boyle (2005)." *Psychological Bull.* 131 (1): 61–65.
Raver, C. Cybele, Stephanie M. Jones, Christine Li-Grining, Fuhua Zhai, Kristen Bub, and Emily Pressler. 2011. "CSRP's Impact on Low-Income Preschoolers' Preacademic Skills: Self-Regulation as a Mediating Mechanism." *Child Development* 82 (1): 362–78.
Rhineland-Palatinate Statistics Office. 2018. *Allgemeinbildende Schulen im Schuljahr 2017/2018.* Bad Ems: Statistisches Landesamt Rheinland-Pfalz.
Riggs, Nathaniel R., Mark T. Greenberg, Carol A. Kusché, and Mary Ann Pentz. 2006. "The Mediational Role of Neurocognition in the Behavioral Outcomes of a Social-Emotional Prevention Program in Elementary School Students: Effects of the Paths Curriculum." *Prevention Sci.* 7 (1): 91–102.
Roberts, Gehan, Jon Quach, Megan Spencer-Smith, Peter J. Anderson, Susan Gathercole, Lisa Gold, Kah-Ling Sia, et al. 2016. "Academic Outcomes 2 Years after Working Memory Training for Children with Low Working Memory: A Randomized Clinical Trial." *JAMA Pediatrics* 170 (5): e154568.
Rode, Catrin, Robby Robson, Andy Purviance, David C. Geary, and Ulrich Mayr. 2014. "Is Working Memory Training Effective? A Study in a School Setting." *PLoS ONE* 9 (8): e104796.
Romano, Joseph P., and Michael Wolf. 2005. "Stepwise Multiple Testing as Formalized Data Snooping." *Econometrica* 73 (4): 1237–82.

———. 2016. “Efficient Computation of Adjusted *p*-Values for Resampling-Based Stepdown Multiple Testing.” *Statis. and Probability Letters* 113:38–40.

Sala, Giovanni, N. Deniz Aksayli, K. Semir Tatlidil, Tomoko Tatsumi, Yaduyuki Gondo, and Fernand Gobet. 2019. “Near and Far Transfer in Cognitive Training: A Second-Order Meta-Analysis.” *Collabra: Psychology* 5 (1): 18.

Sala, Giovanni, and Fernand Gobet. 2020. “Working Memory Training in Typically Developing Children: A Multilevel Meta-Analysis.” *Psychonomic Bull. and Rev.* 27 (3): 423–34. https://doi.org/10.3758/s13423-019-01681-y.

Schmeichel, Brandon J., Rachael N. Volokhov, and Heath A. Dernaree. 2008. “Working Memory Capacity and the Self-Regulation of Emotional Expression and Experience.” *J. Personality and Soc. Psychology* 95 (6): 1526–40.

Schunk, Daniel, Eva M. Berger, Henning Hermes, Kirsten Winkel, and Ernst Fehr. 2022. “Teaching Self-Regulation.” *Nature Human Behaviour* 6 (12): 1680–90.

Shipstead, Zach, Kenny L. Hicks, and Randall W. Engle. 2012. “Cogmed Working Memory Training: Does the Evidence Support the Claims?” *J. Appl. Res. Memory and Cognition* 1 (3): 185–93.

Sisk, Victoria F., Alexander P. Burgoyne, Jingze Sun, Jennifer L. Butler, and Brooke N. Macnamara. 2018. “To What Extent and under Which Circumstances Are Growth Mind-Sets Important to Academic Achievement? Two Meta-Analyses.” *Psychological Sci.* 29 (4): 549–71.

St Clair-Thompson, Helen, Ruth Stevens, Alexandra Hunt, and Emma Bolder. 2010. “Improving Children’s Working Memory and Classroom Performance.” *Educ. Psychology* 30 (2): 203–19.

Van Snellenberg, Jared X., Ragy R. Girgis, Guillermo Horga, Elsmarieke van de Giessen, Mark Slifstein, Najate Ojeil, Jodi J. Weinstein, et al. 2016. “Mechanisms of Working Memory Impairment in Schizophrenia.” *Biological Psychiatry* 80 (8): 617–26.

Walker, Susan P., Susan M. Chang, Amika S. Wright, Rodrigo Pinto, James J. Heckman, and Sally M. Grantham-McGregor. 2022. “Cognitive, Psychosocial, and Behaviour Gains at Age 31 Years from the Jamaica Early Childhood Stimulation Trial.” *J. Child Psychology and Psychiatry* 63 (6): 626–35.

Westerberg, Helena, Tatja Hirvikoski, Hans Forssberg, and Torkel Klingberg. 2004. “Visuo-Spatial Working Memory Span: A Sensitive Measure of Cognitive Deficits in Children with ADHD.” *Child Neuropsychology* 10 (3): 155–61.

Wiley, Jennifer, Andrew F. Jarosz, Patrick J. Cushen, and Gregory J. H. Colflesh. 2011. “New Rule Use Drives the Relation between Working Memory Capacity and Raven’s Advanced Progressive Matrices.” *J. Experimental Psychology: Learning Memory and Cognition* 37 (1): 256–63.

Yeager, David S., Paul Hanselman, Gregory M. Walton, Jared S. Murray, Robert Crosnoe, Chandra Muller, Elizabeth Tipton, et al. 2019. “A National Experiment Reveals Where a Growth Mindset Improves Achievement.” *Nature* 573 (7774): 364–69.

Yeager, David Scott, Rebecca Johnson, Brian James Spitzer, Kali H. Trzesniewski, Joseph Powers, and Carol S. Dweck. 2014. “The Far-Reaching Effects of Believing People Can Change: Implicit Theories of Personality Shape Stress, Health, and Achievement During Adolescence.” *J. Personality and Soc. Psychology* 106 (6): 867–84.