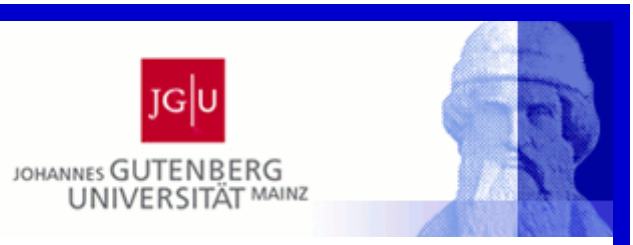


Gutenberg School of Management and Economics

& Research Unit “Interdisciplinary Public Policy”

Discussion Paper Series

Self-productivity and Cross-productivity in the Process of Skill Formation

Eva M. Berger

December 22, 2020

Discussion paper number 2027

Johannes Gutenberg University Mainz
Gutenberg School of Management and Economics
Jakob-Welder-Weg 9
55128 Mainz
Germany
https://wiwi.uni-mainz.de/

Contact Details:

Eva M. Berger
German Council of Economic Experts
Gustav-Stresemann-Ring 11
65189 Wiesbaden, Germany

eva.berger@svr-wirtschaft.de

Self-productivity and Cross-productivity in the Process of Skill Formation

Eva M. Berger*

December 22, 2020

Abstract

Given the insight that individual skills crucially impact various life outcomes, questions about the process of skill formation are increasingly being researched. Evidence about path dependency and about substantial and lasting effects of early childhood events emphasizes the importance of the dynamic component in the skill formation process. This dynamic component has been incorporated in skill formation models featuring self- and cross-productivity, while empirical evidence is scarce. Filling this gap I estimate an instrumental variable model, using as instrument a randomized controlled working memory training intervention, to investigate the question of whether skills are self- and cross-productive, i.e., whether skills boost skills over time. My results show that, first, an exogenous shock to one specific skill (working memory capacity) at an initial stage leads to that same skill being improved in a later stage, but only to the extent of the initial skill shift without any extra effect on the *production* of this skill (self-productivity in the broader sense but not in the narrower sense). Second, I find the exogenously shifted skill, while having no immediate effect on other skills, boosting the production of a number of other skills over time. Hence, I provide evidence about skills being dynamically cross-productive. My findings imply that early disadvantages can be the reason for skill gaps opening up over the life cycle and they explain why early interventions can have significant long-term effects for individual human capital accumulation. My results have implications for the design of policies intended to foster human capital and to augment equality of opportunity.

Keywords: Skill formation, human capital development, dynamic skill production function, self-productivity, cross-productivity, child development, educational intervention

JEL-codes: I21, I24

*German Council of Economic Experts, c/o Statistisches Bundesamt, Gustav-Stresemann-Ring 11, 65189 Wiesbaden, Germany, eva.berger@svr-wirtschaft.de

1 Introduction

It is well documented that various skills (e.g., cognitive skills, social skills, emotional skills) fundamentally determine a range of life outcomes, such as educational attainment, earnings, and health (e.g., Heckman et al. 2006, Lindqvist and Vestman 2011, Heckman 2007, Chiteji 2010, Hanushek et al. 2015, Heckman et al. 2018a,b; for comprehensive surveys see Borghans et al. 2008, Almlund et al. 2011). Given the major importance of skills, a growing amount of research is trying to explain the complex process of skill formation. A deep understanding of the process of skill formation is crucial for understanding the origins of inequality in socio-economic outcomes. It is also fundamental for policy makers trying to design effective and efficient instruments that facilitate skill formation and support disadvantaged groups suffering from skill gaps and thus limited life opportunities.

The timing and the dynamics seem to play a key role in the process of skill formation. Gaps in skills across children open up at early ages—even before schooling begins—and persist or even widen over the life cycle (Heckman and Carneiro 2003, Cunha et al. 2006, Blomeyer et al. 2009). This pattern has been documented for various types of skills, including academic as well as social skills. It explains the observation of growing skill gaps between different socio-economic groups. In addition, evidence shows that early childhood environments importantly shape later life outcomes and that even mild early life shocks can have lasting consequences (Currie and Almond 2011, Almond et al. 2018). Interventions taking place in early stages of childhood have been found to effectively reduce skill gaps (see Heckman and Kautz 2014, for a review of this literature). Later programs, in contrast, seem to be less effective. Overall, the economic returns to early educational interventions tend to exceed those from programs aimed at adolescents or young adults (Heckman 2006, Knudsen et al. 2006). All this evidence points to the importance of timing and dynamic interdependencies in the process of skill formation.

Given that evidence, it is important to improve the understanding about the dynamics in the skill production process. In particular, the question arises to what extent skills at one stage play a role in the formation of skills at later stages. The process of a skill at one stage positively affecting this same skill at a later stage is termed 'self-productivity', following the seminal work by Cunha and Heckman (2007), who propose a dynamic production function of skills. The process of a skill at one stage positively affecting *other* skills at a later stage is termed 'cross-productivity'. Skills being self- and cross-productive could explain why skill gaps between children increase over the life cycle and why schools fail to equalize starting opportunities. Evidence about skills being self- and cross-productive would then corroborate the importance of human capital investments being made early in life. So far, however, comprehensive evidence about skills being self- and cross-productive is missing—given the enormous empirical challenges in its clean analysis.

The present paper aims at filling this gap. I provide estimates of self- and cross-productivity of skills, i.e., I empirically answer the question of whether the stock of a skill in one period

causally improves the stock of this skill (*self-productivity*) and other skills (*cross-productivity*) in subsequent periods. I thus investigate a key aspect in the production technology of skills. I use rich panel data on a variety of primary school students' skills and exploit the exogenous increase in one skill being the result of a specific randomized intervention. My data allow me to estimate in a reduced-form manner the self- and cross-productivity, i.e., the causal effect of one skill on the further development of that same as well as other skills over time.

The randomized intervention I exploit is a computer-based working memory (WM) training, the exogenously manipulated skill thus is WM capacity. Working memory is the capacity to mentally store and process information and plays an important role in many activities. It has been researched on extensively in psychology and neuroscience (Baddeley 1999). In particular, WM has been documented to play an important role in the process of learning (Bergman Nutley and Söderqvist 2017, Holmes et al. 2009) and individuals with learning problems often have low WM capacity (Martinussen et al. 2005). This is why WM capacity is particularly likely to have dynamic effects in the skill formation process.

Furthermore, WM capacity has been documented to be malleable by training (Aksayli et al. 2019, Melby-Lervåg et al. 2016, Sala and Gobet 2020, Shipstead et al. 2012). The results from our intervention documented in Berger et al. (2020) confirm this conclusion. Moreover, normal school routine does not explicitely train WM capacity and the WM training intervention is thus likely to have a measurable impact for a longer period of time. This is in contrast to interventions targeting skills that are focussed on in subsequent human capital investments anyway (Bailey et al. 2020), such as arithmetic skills focussed on by regular school lessons. The effects of a, say, arithmetics intervention, even though potentially being effective in the short-run, are likely to fade out over time because the subsequent regular school lessons work as a substitute for the arithmetics intervention and thus make the control group catch up with the treatment group over time. This is different for the skill our intervention focusses on, WM capacity, which is not the direct focus of regular school lessons. Hence, the intervention meets the criteria outlined by Bailey et al. (2017) for treatment effects being persistent: according to these authors, interventions should target at what they call "trifecta" skills—ones that are malleable, fundamental, and would not have developed in the absence of the intervention.

Identifying skills' self- and cross-productivity is challenging for at least two reasons. The first is the omited variables bias: Factors in the environment of children affect skills at several stages. Skill gaps widening over the life cycle could certainly be the result of skills being self- and cross-productive. However, alternatively, it could be the result of differences in the stimulating environment across children and these different environments persisting and affecting skills at various stages. A variety of factors in a child's environment influence her skill development, and environments vary across children (e.g., across socio-economic groups). The number of factors that influence a child's skill development is enormous. Relevant factors include the number and variety of words spoken at home (Hart and Risley 1995), the amount of time children spend in

educational activities with their parents and parenting style (Fiorini and Keane 2014), neighborhood stability (Gibbons et al. 2017), exposure to disruptive peers in the classroom (Carrell et al. 2018), ordinal academic rank (Murphy and Weinhardt 2020, Elsner and Isphording 2017), and teacher quality (Kane et al. 2011, Jackson 2018)—to name just a few. The examples illustrate that countless and in part hardly measurable factors in a child's environment affect skill development and that it is impossible to fully observe and control all these factors when estimating the skill production function. Estimations of the effect of early skills on later skills will therefore usually suffer from omitted variables bias. In order to get unbiased estimates it is of major importance to exploit an exogenous source of variation in the skill level at the earlier stage in the analysis. This is the fundamental advantage of the data used in this paper: I exploit the exogenous variation in WM capacity being the result of a specific randomized controlled intervention. In a two-stage estimation model I use the treatment indicator as an instrument for WM capacity at the early stage and thus consistently estimate the self- and cross-productivity effect of WM capacity.

The second challenge for empirical studies on skills' self- and cross-productivity is the multiplicity of skills. Besides cognitive ability (typically measured by IQ tests), a number of other skills, such as self-regulatory, social, and emotional skills, have been documented to play an important and independent role for various life outcomes (Almlund et al. 2011, Borghans et al. 2008, Kautz et al. 2014, Bowles et al. 2001, Moffitt et al. 2011, Dohmen et al. 2009, Backes-Gellner et al. 2018). The fact of skills being multiple in nature implies the possibillity of dynamic interdependencies in the skill production process. As an example, one could think of a highly attentive child being able to improve math skills at school more than a child that is less attentive. A child with a strong memory might improve her language skills faster than a child with a weaker memory. Emotional skills (emotional security) fosters child exploration and more vigorous learning of cognitive skills (Cunha and Heckman 2007). Hence, apart from self-productivity (i.e., the stock of a skill in one period shapes *this same* skill in future periods), cross-productivity (i.e., the stock of a skill in one period shapes *other* skills in future periods) has to be taken into account when exploring the dynamic skill formation process. The challenge in identifying self- and cross-productivity of some skill j lies not only in the requirement of a randomized intervention that generates an exogenous variation of skill j, it lies also in the requirement of the intervention being *specific* in the sense that it directly changes *only* skill j without directly changing other skills nor environments. This is necessary in order to be able to causally trace changes in skills at later stages back to changes in skill j rather than to other skills or environments at the earlier stage.

The intervention I rely on, the WM training intervention, is very specifically targeted to only WM capacity. Direct effects of the training on other skills are close to zero (see Section 5 below as well as the intervention results analyzed in detail in Berger et al. (2020)). This is consistent with the evidence privided in the literature documenting a zero short-term effect of WM training on skills other than WM capacity (for reviews see, e.g., Aksayli et al. (2019), Melby-Lervåg et al.

(2016), and Shipstead et al. (2012)). Also, given that the WM training in our field experiment (Berger et al. 2020) was integrated in the normal school routine—similar to any other sequence of exercises introduced to children during a school year—the intervention did not change other aspects of the school context nor the home environment. Parents' consent was not required for the training but only for the data collection—both for the treatment and the control group. This educational intervention is thus special in terms of its specificity. In this respect the intervention differs from other recent educational interventions, which involve a variety of changes in schools or home environments and thus are highly effective in directly impacting various skills.[1] In contrast, the intervention this paper relies on is specifically targeted at WM capacity only, and thus it is uniquely suitable to identify skills' self- and cross-productivity, a key feature of the skill production function.

The key results of this paper are as follows: First, I find that the target skill (WM capacity) at one stage has a positive effect on the same skill measured at later stages. This means that I find evidence for WM capacity being self-productive in the broader sense. But this self-productivity effect appears to be close to one, i.e., it consists of a pure *level* effect with no additional effect on skill *production*. In other words, the WM training appears to have shifted the level of WM capacity and this higher level remains stable over time, i.e., neither it fades out nor does the growth path of WM capacity become steaper. The latter, however, would be expected if one hypothesized WM capacity to have an effect on own production. Second, with respect to cross-productivity, I do find an effect of WM capacity on the subsequent *production* of skills other than WM capacity: I find WM capacity to be cross-productive for geometry skills, the ability to inhibit pre-potent impulses, and fluid IQ (measured by Raven's matrices). Hence, I find that WM capacity plays a role in the development of other skills. Showing that the stock of a skill can affect the growth path of other skills, I reveal an important feature in the skill formation process.

It would be unreasonable to assume that the extent of self- or cross-productivity effects be equal across all skills. Possibly, basic skills such as memory or reasoning capacities affect the development of other skills to a greater extent than specific/applied skills such as calculation skills

[1] An example of a recent educational intervention includes the PATH ('Promoting Alternative Thinking Strategies') program, which is an intensive one-year, teacher-run training program, which also involves parents, and which targets socio-emotional skills such as self-control, empathy, emotional literacy, and interpersonal problem-solving skills of eight year old children in Switzerland (Sorrenti et al. 2020). Another recent example is the one-year mentoring program provided to second-grade primary school students in Bonn/Germany (Kosse et al. 2020). The grit intervention conducted with elementary school students in Istanbul (Alan et al. 2019, p. 1128) involves a curriculum designed to "highlight (i) the plasticity of the human brain against the notion of innate ability, (ii) the role of effort in enhancing skills and achieving goals, (iii) the importance of a constructive interpretation of setbacks and failures, and (iv) the importance of goal setting". The reading intervention with second-grade children in Aarhus/Denmark consisted of a growth mindset approach to parents combined with delivery of books and encouragement to read together with the child Andersen and Nielsen (2016). Older and well-known programs include the Perry Preschool Program and the Abecedarian Program, which both comprise intensive child care programs and home visits Cunha et al. (2006). All the mentioned interventions encompass a variety of interventional measures and skills targeted. This was reasonable given that the aim of those interventions was to be as effective as possible in improving important skills and later life outcomes. The intervention analyzed in the present paper is different in that it specifically targets WM capacity only.

or foreign language skills do. Therefore, I certainly refrain from generalizing my findings by claiming that *all* skills be cross-productive. Also, it is likely that the cross-productivity potential varies not only across trigger skills but also across 'cross-skills' (other skills). In this study, I found WM capacity to be cross-productive for geometry skills but not for arithmetic skills, for Raven's IQ but not for reading skills. The pattern of results might be different when investigating a different trigger skill instead of WM capacity. Nevertheless, my results are crucial for research and policy advice: I document that cross-productivity plays a role for at least some skills, and therefore identifying skills that are cross-productive and fostering these skills early in the life cycle can importantly affect human capital development.

My research relates to previous studies that aimed at estimating the dynamic production function of skills, in particular to the seminal articles of Cunha and Heckman (2008) as well as Cunha et al. (2010). The authors provide a comprehensive model of skill production where skills and investments are dynamically interrelated across stages in childhood. They assume skills to reduce to two latent factors, one being measured by a mathematics and a reading recognition test (from PIAT) and the other being measured by a behavioral problem index. I contribute to the literature, first, by allowing for a finer differentiation between skills, both among the cognitive as well as among the noncognitive skills. Having a great battery of skills, all measured repeatedly by objective and highly standardized tests, I address the challenge of the dynamic interrelation of multiple skills. Second, the empirical approach by Cunha and Heckman (2008) and Cunha et al. (2010) is different from mine: the authors estimate a structural model by explicitely including parental factors that affect skill formation—termed 'investments' into skills. In contrast, I exploit the exogenous variation of one skill level being the result of a randomized controlled intervention study. I am thus able to estimate self- and cross-productivity in skill formation by a reduced-form approach and refrain from structural modeling assumptions.

The present paper contributes to the understanding of the dynamic process of skill formation. This is crucial for understanding the sustainability of inequalities in skills and life outcomes within the population of many countries. Skills being self- and cross-productive explains (part of) the observed path dependency in skill formation. It also contributes to understanding the limited intergenerational mobility in socio-economic status observed in many countries. Furthermore, an improved understanding of the dynamic skill formation process is also needed for optimally designing policies aimed at fostering human capital development. Policy makers have to decide about the skills targeted by interventions, the timing of intervention, and the amount spent for interventions. My findings suggest to focus on skills that are cross-productive and to focus on early stages in the children's life cycle. Due to the dynamics in the skill production, the benefits of educational interventions can be very high in future periods and the returns to educational expenditure are thus larger than what is concluded when considering only immediate educational improvements. This implies that the optimal amount spent for educational policies is higher than what would be inferred from considerations of only short-term educational effects.

The rest of the paper is organized as follows: In Section 2, I formalize the development of skills in a dynamic model that accounts for multiple skill dimensions. In Section 3, I describe the empirical strategy employed to estimate self- and cross-productivity. Section 4 presents the experimental design and data collection. Section 5 presents and discusses the results. Section 6 concludes.

2 The Model of Skill Formation

The analyses in this paper are based on a dynamic model of skill formation. The stock of a child's skills in period $t + 1$, θ_{t+1}, is a function of the stock of skills in the previous period θ_t and initial environmental factors X:

$$\theta_{t+1} = f(\theta_t, X) \tag{1}$$

The vector of environmental factors X includes all initial conditions that are exogenous to the child's skills but affect the production of the child's skills.

Without loss of generality, the skill formation function $f(.)$ can be split into two components: the previous stock of skills θ_t and the change in skills $\Delta\theta_t$. The change in skills is a function of θ_t and X, such that

$$\theta_{t+1} = \theta_t + g(\theta_t, X). \tag{2}$$

Equation (2) can be interpreted as follows: The stock of skills in period $t + 1$ is equal to the sum of the stock of skills from the previous period and the skill production since the last period. The production component can be positive (learning) or negative (depreciation of skills over time, i.e., unlearning).

In the next step I incorporate the fact that skills are not one-dimensional but have many facets. A child can have high mathematical skills and weak verbal skills, while another child might have high verbal skills and low mathematical skills. Even for skills that are contemporaneously orthogonal, a dynamic interdependency might exist. Such a dynamic interdependency emerges for example if a child with strong verbal skills has an advantage in reading and understanding mathematical exercises and explanations and therefore improves mathematical skills faster than a child with weak verbal skills. It is therefore important to, first, allow for skills to be multidimensional and, second, allow for their dynamic interdependency. I therefore expand the model as follows: I consider two contemporaneously orthogonal skills, skill j and skill k ($j \neq k$) and allow the production of each skill to depend on the stock of *both* skills in the previous period. The stock of skill j in period $t + 1$ thus is

$$\theta^j_{t+1} = \theta^j_t + g_j(\theta^j_t, \theta^k_t, X),$$

and the stock of skill k in period $t+1$ is

$$\theta^k_{t+1} = \theta^k_t + g_k(\theta^j_t, \theta^k_t, X).$$

I call the partial derivative of θ^j_{t+1} with respect to θ^j_t "self-productivity". It is given by

$$\frac{\partial \theta^j_{t+1}}{\partial \theta^j_t} = 1 + \frac{\partial g_j(\theta^j_t, \theta^k_t, X)}{\partial \theta^j_t}. \tag{3}$$

It results to be the sum of the level effect (which trivially is equal to one) and the production effect (which is $\frac{\partial g_j(\theta^j_t, \theta^k_t, X)}{\partial \theta^j_t}$).

I call the partial derivative of θ^k_{t+1} with respect to θ^j_t "cross-productivity". It is given by

$$\frac{\partial \theta^k_{t+1}}{\partial \theta^j_t} = 0 + \frac{\partial g_k(\theta^j_t, \theta^k_t, X)}{\partial \theta^j_t}. \tag{4}$$

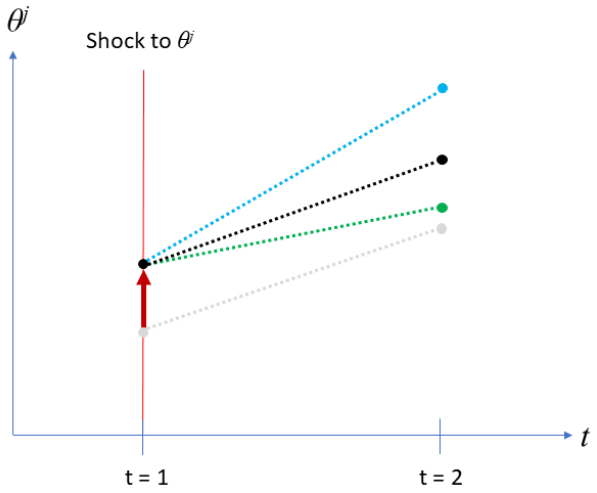
In equation (4) the level effect is zero and the cross-productivity effect thus reduces to the production effect.

Three different scenarios for the self-productivity, i.e., for the development of skill j after an exogenous shock to skill j, are illustrated in Figure 1a. The scenarios vary in the value of the production effect, i.e., in the sign of the second term of the right hand side of equation (3). The red vertical line in the figure marks the period in which skill j is exogenously increased ($t = 1$). The black dot in the subsequent period ($t = 2$) marks the level of skill j for the scenario in which the production effect is zero. In this scenario the difference between the actual skill level and the counterfactual skill level (grey dot) in $t = 2$ is equal to the difference between actual and counterfactual skill levels in $t = 1$. I.e., the self-productivity is one (due to the first term of the right hand side in equation (3)). The blue dot and dotted line, in contrast, illustrate the skill development in the presence of a *positive* production effect. In this scenario, the difference between the actual and the counterfactual skill level is *larger* in $t = 2$ than in $t = 1$. I.e., the self-productivity is larger than one. The green dot (and dotted line) illustrates the skill development in the presence of a *negative* production effect: the difference between the actual and the counterfactual skill level is *smaller* in $t = 2$ than in $t = 1$. I.e., the self-productivity is smaller than one.

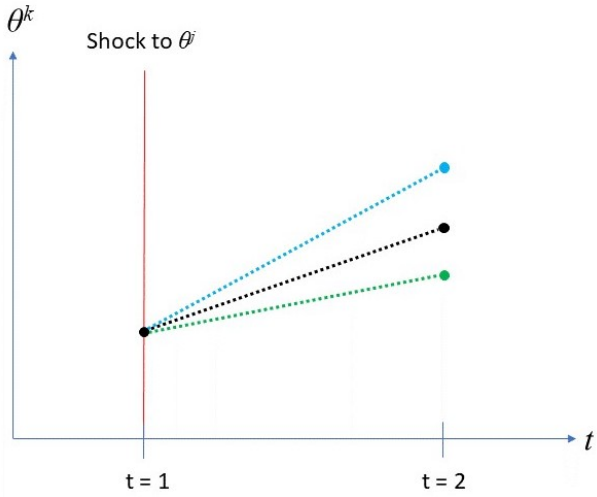
Analogously, different scenarios for the cross-productivity, i.e., for the development of skill k after an exogenous shift of skill j are illustrated in Figure 1b. Again, the red vertical line in the figure marks the period ($t = 1$) in which skill j is exogenously increased, while the vertical axis measures the level of skill k. The level of skill k in period $t = 1$ is unaffected by the exogenous shock to skill j, this is a consequence of the two skills being contemporaneously orthogonal. The black dot in $t = 2$ marks the level of skill k in the case that the production effect is zero: the level of skill k is equal to the counterfactual skill level. This is the case if the cross-productivity is

Figure 1: Self- and cross-productivity: The effect of increasing the level of skill j on the subsequent development of skill j and the contemporaneously orthogonal skill k, three scenarios each

(a) Self-productivity: The effect on skill j



(b) Cross-productivity: The effect on skill k



zero due to the production effect being zero (second term of the right hand side of equation (4)). In this case, skill j being shifted has no consequences for the development of skill k. The blue dot, in contrast, illustrates the development of skill k in the presence of a *positive* production effect and thus a positive cross-productivity: the actual level of skill k in $t = 2$ is *larger* than the counterfactual level of skill k. The green dot illustrates the development of skill k in the presence of a *negative* production effect and thus a negative cross-productivity: the actual level of skill k in $t = 2$ is *smaller* than the counterfactual level of skill k in that period. To sum up, given that the two skills are contemporaneously orthogonal, the exogenously shifted skill j can affect the stock of skill k only through an effect on its subsequent production.

In my empirical analysis I will shed light on the sign of the production effect in both self- and cross-productivity by comparing the empirical skill profile to the different theretical scenarios.

3 Empirical Strategy

My empirical strategy to identify self-productivity, i.e., the effect of the current stock of skill j (θ_t^j) on the future stock of that same skill j (θ_{t+1}^j), and cross-productivity, i.e., the effect of the current stock of skill j on the future stock of a different skill k (θ_{t+1}^k), relies on exploiting an exogenous shift of the stock of skill j. Skill j in my application is WM capacity and the source of exogenous variation is a randomized intervention that integrated a 5-week computer-based WM training in primary schools. Children's skills are measured before the training, shortly after the training as well as 6 and 12–13 months after the training.

My empirical model is summarized by the following two equations:

$$\theta_{i,t+1}^k = \alpha_0^k + \alpha_1^k \theta_{i,t}^{\text{WMC}} + \alpha_2^k X_i + \varepsilon_{i,t+1}^k \tag{5}$$

$$\theta_{i,t}^{\text{WMC}} = \beta_0 + \beta_1 \text{WMT}_i + \beta_2 X_i + \eta_{i,t+1} \tag{6}$$

WMT_i is a binary variable indicating whether child i has been assigned to the treatment condition, X_i is a vector of exogenous environmental factors relevant to skill formation, and WMC indicates working memory capacity. I estimate the model using Two Stage Least Squares (2SLS) so that equation (6) is the first stage and WMT_i serves as an instrumental variable for $\theta_{i,t}^{\text{WMC}}$. Replacing k by WMC in equation (5), the parameter α_1 will be the self-productivity parameter to be estimated. Replacing k by a skill other than WM capacity, α_1 represents the respective cross-productivity parameter to be estimated.

The self-productivity effect will be equal to one if the production effect (second term of the right hand side of equation (3)) is zero, larger than one if the production effect is positive, and smaller than one if the production effect is negative. The cross-productivity effects, in contrast, will be estimated to be zero if the production effect (second term of the right hand side of equa-

tion (4)) is zero, larger than zero if the production effect is positive, and smaller than zero if the production effect is negative.

All skill measures that I use in my estimations are scores standardized within the control group of each evaluation wave to mean = 0 and standard deviation = 1. Hence, estimated effect sizes can be interpreted in terms of fractions of a standard deviation. As outlined, instead of using the skill *change* as dependent variable, I estimate the effect on skill *levels* and control for the pre-treatment level of the respective skill. The advantage of this method is that the variance of the estimated effect is smaller, i.e., the effect is measured with higher precision (McKenzie 2012, Frison and Pocock 1992). All models further include school fixed effects as well as controls for gender and age. Finally, in order to account for dependencies of observations within school classes, standard errors are clustered at the classroom level.

4 Data

The data analyzed in this paper come from a randomized-controlled intervention study that was conducted in primary schools in Mainz, Germany, in 2013/2014 (see Berger et al. 2020).

4.1 Participants

After having received ethical approval in September 2012, 31 first grade classes were recruited 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, 580 parents provided their consent to the data collection (consent rate of 96.8%). Test data of the relevant outcome measures in all four evaluation waves, i.e., prior to the treatment (W1), shortly after the treatment (W2), 5–6 months after the treatment (W3) and 12–13 months after the treatment (W4), were successfully collected for 518 children. Attrition over the course of the study (from W1 to W4) was very low (7%) with no difference between treatment and control group (see Online Appendix Section A.1 for details).

4.2 Treatment and Control Condition

Randomization into treatment and control group was realized between classes and within schools: 15 classes (249 children, i.e., 48%) were randomly assigned to the treatment group and 16 classes (269 children) to the control group. 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. It was introduced to the children as a normal sequence of exercises, similar to when the teacher introduces a new sequence of exercises for math, reading, or writing required

by the curriculum. Children could not opt out of the WM training and no parental consensus was required as the training was integrated into the usual classroom activities. The usual teacher supervised the WM training lessons and children remained in their classroom and conducted the training sessions at their usual desks.

The training was conducted with a commercially available WM training software[2] providing training on different span tasks, using an age-specific user-interface. As an example, in one of the tasks (called 'Rotating data link') a panel of 16 lamps arranged in a 4x4 grid is shown. Lamps light up in a certain order; then the panel turns by 45 degree; and finally the child has to click on the lamps in the right order. In another task (called 'Asteroids') asteroids light up while floating through the space and children must click on them in the right order. In the task called 'Input module' buttons with the numbers 1 to 9 arranged on a 3x3 grid are shown. Some numbers are read out loudly while lighting up; subsequently, the child has to click on the buttons in reverse order. In total, children are trained on ten different tasks. Eight focus on purely visuo-spatial WM, while two include elements of verbal WM training. The exercises become progressively more challenging over the course of the training and adapt to the individual capacity level of each child.

The training software is very specifically targeted to improve WM capacity. It is unlikely that it directly improves other (e.g., academic) skills. However, it might well improve the *acquisition process* for these other skills. This is what I examine in the empirical part of this paper.

Logins for the training software were user-specific and only valid during the intervention period. The children thus had access to the training software only during their dedicated training sessions. Hence, spill-over effects to the control group are impossible.

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

The WM training in this study 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. Given that the curriculum, which had to be covered until the end of the school year, remained unchanged, the WM training lessons essentially replaced practice lessons in math and German.

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

4.3 Data Collection

A professional data collection service provider ran the four evaluation waves: prior to the treatment (W1), shortly after the treatment (W2), 5–6 months after the treatment (W3), and 12–13 months after the treatment (W4) (for further details, see Online Appendix Section A.3).

In each evaluation wave, the children completed highly standardized tests administered by staff that was blind to treatment conditions. Parents of both treatment and control children gave their consent to participate in the data collection (consent rate of 96.8%). 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. The entire sequence of tests was computer-based, including auditive (via headphones) explanations and comprehension checks. The input devices for the evaluation tasks were not computer mice but large touchscreens in order to avoid any bias arising from the fact that children in the treatment group had been working with computer mice during the WM training. When the children had finished all evaluation tasks in a given wave, they were rewarded for their participation with a selection of toys to ensure high motivation. Test and rewards were the same for children in control and treatment condition.

Each data collection wave comprises tests on the following skill dimensions: WM capacity, geometry, arithmetic, reading, fluid IQ, inhibition control, and sustained attention. The structure of each test was similar across waves, but the difficulty level of items was adjusted according to the children's development over time. A pretest prior to W1 with a different (smaller) sample of similar aged children served to adapt the initial level of difficulty.

Working Memory Capacity Measure

WM capacity was measured by a visuo-spatial complex span task. The task consisted of a series of screens each of which showed three symbols arranged in a row. The child had to identify the slightly deviant symbol within the three in each screen. After a sequence of screens, the child had to recall the position of the deviant symbols in the correct order. The task clearly differs from all WM training tasks. In addition to the visuo-spatial WM task, two verbal WM tasks have been administered to the children in the course of the study. Since the training, however, focuses on visuo-spatial WM, I also focus on the visuo-spatial WM measure when analyzing self- and cross-productivity effects in this paper.

Educational Skill Measures

Educational achievement was measured in three areas: arithmetic, geometry, and reading. Skills in geometry were used as an outcome measure because—like arithmetic and reading—it plays an important role in everyday life (e.g., orientation, reading maps, driving) as well as in various professions (e.g., construction/architecture, fashion/art design, geography, astronomy, physics,

sports, etc.) and therefore is an important part of the math curriculum at school. (For further details on the educational achievement tests, see Online Appendix Section A.4.)

Other Skill Measures

In addition, three other tests were administered to the children; the tests 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 Coloured Progressive Matrices test (Bulheller and Häcker 2010), the ability to inhibit pre-potent responses was measured by the go/no-go task (Gawrilow and Gollwitzer 2008), and attentional stamina was measured by the bp task (Esser et al. 2008). (For further details on these tests, see Online Appendix Section A.4).)

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 A12). 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 et al. 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 A13). 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.

Questionnaires

In addition to the skill testing of children, questionnaires were addressed to parents and teachers. Parent questionnaires were only distributed in the data collection waves W1 and W3, i.e., before the intervention and 5–6 months after the intervention. Parent questionnaires included a number of questions on background characteristics that we use to check sample balance. Parents' response rate was 82% in W1 and 77% in W3. Teachers filled out a questionnaire in each data collection wave. The questionnaire contained questions on children's characteristics—such as

their migration background or language problems—and teacher characteristics. The return rate for the teacher questionnaire was 100% in all four data collection waves.

4.4 Summary Statistics

Summary statistics of the relevant variables are reported in Table 1. About 50% 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, standard deviation = 4.4 months). Gender and age variables (including age at test days) are taken from parental consent forms and are therefore available for all children. The variables migration background and language problems stem from the teacher questionnaire administered in W1, the variables on income and mother's educational degree stem from the parental questionnaire in W1.

Table 1: Summary Statistics

Variable	Mean	Std. Dev.	N
Treatment	0.481	0.500	518
Male	0.496	0.500	518
Age in months on Jan 1, 2013	82.205	4.356	518
Age on test day w1 (in months)	84.335	4.406	518
Age on test day w2 (in months)	87.377	4.384	518
Age on test day w3 (in months)	92.464	4.394	518
Age on test day w4 (in months)	99.622	4.391	518
Migration background	0.447	0.498	514
Language problems	0.241	0.428	518
Monthly HH Net Income <750 Euros	0.017	0.131	402
Monthly HH Net Income 750-1500 Euros	0.109	0.313	402
Monthly HH Net Income 1500-2500 Euros	0.211	0.409	402
Monthly HH Net Income 2500-5000 Euros	0.435	0.496	402
Monthly HH Net Income >5000 Euros	0.226	0.419	402
Mother university degree	0.460	0.499	404
Mother vocational degree	0.416	0.493	404
Mother no professional degree	0.124	0.330	404

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 (prior to the intervention); 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.

5 Empirical Results

5.1 Sample Balance

To examine whether randomization in the study led to a balanced sample across treatment and control group in terms of socio-economic characteristics and outcome measures, I regress various socio-demographic characteristics (gender, age, migration background, parental income and parental education) measured prior to the treatment (W1) on the treatment indicator. In addition, I test for differences between treatment and control group in the test performance prior to treatment (W1). The results reported in Tables 2 and 3[3] show no significant imbalances between treatment and control group prior to the intervention.

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

	(1) Male	(2) Age on Jan 1, 2013	(3) Migration background	(4) Language problems	(5) Low income (< Eur2500)	(6) Mother univ degr
Treatment	-0.037	0.206	-0.013	0.063	-0.018	-0.001
	(0.032)	(0.357)	(0.056)	(0.048)	(0.049)	(0.044)
N	518	518	514	518	402	404

The results are based on least squares models including school fixed effects. Standard errors in parentheses are clustered at the classroom level. * p<0.05. 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 Pre-Treatment (W1) Test Scores on the Treatment Indicator

	(1) WMC	(2) Geometry	(3) Arithmetic	(4) Reading	(5) Raven's IQ	(6) Inhibition control	(7) Sustained attention
Treatment	-0.070	0.083	0.102	0.093	0.051	-0.105	0.137
	(0.102)	(0.087)	(0.072)	(0.124)	(0.063)	(0.092)	(0.108)
N	515	515	500	514	515	514	513

The results are based on least squares models including school fixed effects as well as gender and age controls. All outcome scores are standardized to mean = 0 and SD = 1. Standard errors in parentheses are clustered at the classroom level. * p<0.05.

[3]The sample size slightly varies across outcomes due to technical reasons during the testing procedure. This is not related to the children's skills nor to treatment assignment. Restricting the sample to only those children for whom all test scores are available does not alter my results.

5.2 Direct effect of treatment on outcomes

In order to be able to estimate the 2SLS model presented in Section 3 the treatment needs to sufficiently strongly affect WM capacity without affecting other skills. Table 4 reports results from regressions of the standardized skill scores measured shortly after the training on the treatment indicator (for a similar estimation, see Berger et al. (2020)). The models include school fixed effects as well as controls for gender, age, and the respective pre-treatment skill level. The treatment effect on WM capacity is sizeable at 0.345, which means that the WM training has increased WM capacity by 34.5% of a standard deviation. The effect is significant at p< 0.001, the F-statistic is with 25 well above the rule of thump of 10 and I thus conclude that it can serve as a sufficiently strong first stage.

Table 4: Effect of treatment on skills measured shortly after the intervention (W2)

	(1) WMC	(2) Geometry	(3) Arithmetic	(4) Reading	(5) Raven's IQ	(6) Inhibition control	(7) Sustained attention
Treatment	0.345*	0.117	-0.022	-0.060	0.056	-0.230*	0.046
	(0.069)	(0.063)	(0.077)	(0.062)	(0.067)	(0.067)	(0.046)
N	515	515	499	512	514	513	513

The results are based on least squares models including school fixed effects as well as controls for gender, age, and the respective pre-treatment skill level. Skill variables are standardized to mean = 0 and SD = 1. Standard errors in parentheses are clustered at the classroom level. * p<0.05.

The estimated treatment effects on all other skill dimensions is insignificant except for inhibition control, where the coefficient appears to be negative (see columns 2–7 of Table 4). These results suggest that the treatment did not directly improve skills other than WM capacity. These findings thus confirm expectations based on the fact that the WM training is very specifically focussing on WM capacity without including any elements close to math or reading tasks for example. The findings are also consistent to what is known from earlier research on WM training, namely that WM training is effective in improving WM capacity but unlikely to directly improve other skills (for reviews of this literature see Aksayli et al. (2019), Melby-Lervåg et al. (2016), and Shipstead et al. (2012)). Certainly, it is hard to prove a null effect; but I will provide further evidence in Section 5.3) below showing that it is unlikely that the 2SLS results below are driven by a direct effect of the treatment on skills other than WM capacity.

5.3 Self- and Cross-Productivity of WM Capacity

The 2SLS estimation results of model (5)–(6) are presented in Table 5. Column 1 reports the self-productivity effect, i.e., the effect of WM capacity at period W2 (period t in the model) on WM capacity at period W4 (period $t + 1$ in the model). It turns out to be significantly positive, the point estimate being 1.033, i.e., close to one and not significantly different from one. Thus

the hypothesis cannot be reject that the self-productivity effect of WM capacity reduces to the level effect (first term on the right hand side of equation (3)) and the production effect (second term on the right hand side of equation (3)) is zero. This implies that improving WM capacity at some point in time has a positive effect on WM capacity at later periods, but only to the same extent as the initial improvement, without an effect on the growth path of WM capacity.

Table 5: 2SLS estimates of self- and cross-productivity

	(1) WMC	(2) Geometry	(3) Arithmetic	(4) Reading	(5) Raven's IQ	(6) Inhibition control	(7) Sustained attention
WMC W2	1.033*	1.000*	-0.283	0.006	0.444*	0.409*	0.551
	(0.274)	(0.335)	(0.274)	(0.360)	(0.187)	(0.192)	(0.352)
N	515	514	500	514	515	514	512

The results are based on two stage least square models using skill levels measured in W4 as dependent variables and the treatment indicator as instrument for working memory capacity (WMC) in W2. The models further include school fixed effects as well as controls for gender, age, and the respective pre-treatment skill level. The skill variables are standardized to mean = 0 and standard deviation = 1. Standard errors in parentheses are clustered on the classroom level. * p<0.05.

Columns 2, 3, and 4 of Table 5 show the cross-productivity effects of WM capacity on educational skills, i.e., on geometry, arithmetic, and reading. For geometry the effect is significantly positive of size 1. Given that in cross-productivity effects the level effect is always zero for contemporaneously orthogonal skills, the estimate implies that WM capacity has improved the *production* of geometry skills. In contrast, I do not find such a production effect for arithmetic nor for reading skills. Two reasons that could explain this pattern of findings are as follows: First, the area of WM improved by the exogenous treatment is the visuo-spatial WM rather than the verbal WM. And the visuo-spatial WM is naturally closer to geometry skills than to arithmetic or reading skills. Thus it is also more likely to play a role in the production process of geometry skills than in that of arithmetic and reading skills. Second, arithmetic and reading skills are trained intensively in primary schools. This intensive training on average strongly improves children's skills in these areas, but it also generates substantial heterogeneity across children due to heterogenous motivation, effort, and learning skills for example. The additional effect resulting from the exogenously improved WM capacity might thus play a comparatively weaker role for the production of frequently trained skills (arithmetic and reading) than for less frequently trained skills (geometry).

Columns 5, 6, and 7 report cross-productivity effects on Raven's IQ, inhibition control, and sustained attention. The effects on Raven's IQ and inhibition control[4] are significantly positive and sizeable with 0.444 and 0.409, respectively. Again, under the assumption that these skills are contemporaneously orthogonal to WM capacity and the treatment did not directly improve

[4]Using the d'-score based on the go/no-go task instead of the inhibition control score produces similar results, see Online Appendix Section A.4 and Table **??**.

them, the finding implies that the enhancement of WM capacity has improved the *production* of these skills. The effect on sustained attention is not significantly different from zero.

To sum up, I have found that the growth paths of some (though not all) skills depend on the initial stock of WM capacity. As mentioned above, the interpretation is correct only if these skills are contemporaneously orthogonal to WM capacity and if the treatment did not directly improve them. In the following, I provide various checks on these issues.

5.4 Robustness Tests

If the exogenously improved WM capacity has made the growth path of skill development steeper, one should see an effect on the difference between the skill score in a later (W4) and the skill score in an earlier (W2) period. Combining two variables measured with some error, however, boosts the attenuation bias generated by measurement error. A more efficient way of estimation is to include the respective skill score from W2 into the model as a covariate (with the W4 outcome as dependent variable). Results from these modified estimations are reported in Table 6. The findings are very similar to my earlier estimates. I thus conclude that the estimated cross-productivity effects are not driven by direct treatment effects on cross-skills. If that was the case, treatment effects should emerge already in W2 (which they do not as shown in Table 4) and including the W2-score as covariates in the 2SLS estimations (Table 6) should make disappear the cross-productivity effects.

Table 6: 2SLS estimates of cross-productivity—including the respective W2 skill score as covariate

	(1) Geometry	(2) Arithmetic	(3) Reading	(4) Raven's IQ	(5) Inhibition control	(6) Sustained attention
N	514	499	512	514	513	512

The results are based on two stage least square models using skill levels measured in W4 as dependent variables and the treatment indicator as instrument for working memory capacity (WMC) in W2. The models further include school fixed effects as well as controls for gender, age, and the respective pre-treatment (W1) and post-treatment (W2) skill level. The skill variables are standardized to mean = 0 and standard deviation = 1. Standard errors in parentheses are clustered on the classroom level. * $p<0.05$.

As mentioned in Section 3, in all my estimations I include the respective skill scores measured in W1. This is consistent with the model in Section 2. Also, this should increase the precision of the estimates without biasing them because the evaluation wave W1 took place prior to the training intervention. If, however, the sample was not perfectly balanced and the initial level of a skill affected its growth path, my results could be driven by the inclusion of W1 skill scores. Therefore, as a robustness test, I exclude the respective W1-scores. The results are shown in

Table 7. The results do not alter compared to my main specification and I thus conclude that my findings are robust.

Table 7: 2SLS estimates of self- and cross-productivity—excluding the respective W1 score as covariate

	(1) WMC	(2) Geometry	(3) Arithmetic	(4) Reading	(5) Raven's IQ	(6) Inhibition control	(7) Sustained attention
WMC W2	0.981*	0.989*	-0.012	0.154	0.493*	0.386*	0.610
	(0.297)	(0.283)	(0.228)	(0.254)	(0.186)	(0.174)	(0.321)
N	518	517	518	518	518	518	517

The results are based on two stage least square models using skill levels measured in W4 as dependent variables and the treatment indicator as instrument for WMC W2. The models further include school fixed effects as well as controls for gender and age. The skill variables are standardized to mean = 0 and SD = 1. Standard errors in parentheses are clustered on the classroom level. * p<0.05.

I exploit the randomized intervention as an exogenous source of variation in WM capacity measured in W2. If, however, the sample was initially not perfectly balanced with respect to this key skill, the results could be biased. As a robustness test I therefore include the pre-treatment level of a WM capacity (i.e., measured in W1) in the estimations as control variables (see Table 8). The results are very similar to the baseline specification.

Table 8: 2SLS estimates of cross-productivity—including WMC W1 as covariate

	(1) Geometry	(2) Arithmetic	(3) Reading	(4) Raven's IQ	(5) Inhibition control	(6) Sustained attention
WMC W2	0.916*	-0.172	0.042	0.418*	0.425*	0.527
	(0.264)	(0.211)	(0.320)	(0.158)	(0.183)	(0.313)
N	514	500	511	512	511	512

The results are based on two stage least square models using skill levels measured in W4 as dependent variables and the treatment indicator as instrument for working memory capacity (WMC) in W2. The models further include school fixed effects as well as controls for gender, age, and the respective pre-treatment skill level and pre-treatment WMC. The skill variables are standardized to mean = 0 and standard deviation = 1. Standard errors in parentheses are clustered on the classroom level. * p<0.05.

If the different skill measures I use to estimate cross-productivity are not sufficiently contemporaneously orthogonal and the intervention affects at least one of the skills directly, my findings cannot be interpreted as evidence for cross-productivity. To check for this possibility, I include all skill scores of evaluation wave W2 as covariates into the estimation model. The results are shown in Table 9. The estimated self- and cross-productivity effects are unaltered compared to the main specification. This suggests that non-orthogonal skills and direct intervention effects are not an explanation for my findings.

Table 9: 2SLS estimates of self- and cross-productivity—including all W1 skill scores as covariate

	(1) Geometry	(2) Arithmetic	(3) Reading	(4) Raven's IQ	(5) Inhibition control	(6) Sustained attention
WMC W2	1.032*	-0.277	-0.007	0.457*	0.427	0.592
	(0.307)	(0.270)	(0.353)	(0.192)	(0.246)	(0.412)
N	497	498	498	498	498	497

The results are based on two stage least square models using skill levels measured in W4 as dependent variables and the treatment indicator as instrument for working memory capacity (WMC) in W2. The models further include school fixed effects as well as controls for gender, age, and all pre-treatment skill levels. The skill variables are standardized to mean = 0 and standard deviation = 1. Standard errors in parentheses are clustered on the classroom level. * p<0.05.

My main estimations use as dependent variables the skills measured 12–13 months after the intervention (W4). Yet, as mentioned above in Section 4, skills have been measured one more time in between, namely 5–6 months after the intervention (W3). In order to be able to detect self- and cross-productivity effects, a sizeable distance between the skill measurements, i.e., between the exougenously improved WM capacity in W2 and later skill measurements is needed. This is why I focused on skills measured in W4 (12–13 months after the treatment) as the dependent variables. In addition, I estimated the same models using as dependent variables the skills measured in W3 (5–6 months after the treatment). The results are reported in Table A2 in the Online Appendix. The point estimates are similar in sign but most cross-productivity effects are smaller in size and insignificant. This is consistent with the model telling that cross-productivity effects exclusively rely on the production effect and no level effect. Given that the time span between W2 and W3 is substantially shorter (only around five months) than the time span between W2 and W4 (around twelve months), there was substantially less time (around 60% less) for the production of skills until W3 than for the production of skills until W4.

5.5 Discussion of Mechanisms

I have documented a positive cross-production effect implying that an increase in WM capacity positively affects the production of other skills. Relating back to the model and linking it to the model proposed by Cunha and Heckman (2007) as well as by Falk et al. (2020), the question about the mechanism of this effect might arise. One could think of two possible mechanisms, the first through an increased quantity (quality) of investments and the second through and increased productivity of given investments (dynamic complementarity). To model these mechanisms I have to modify the skill production function (equation (1)) by explicitely including human capital investments as an argument. The stock of a child's skills in period $t+1$ is then a function of three

arguments: the stock of skills in the previous period θ_t, human capital investments I_t, and initial environmental factors X. The function then reads as

$$\theta_{t+1} = f_1(\theta_t, I_t, X). \tag{7}$$

As before, the vector of environmental factors X includes all initial conditions that are exogenous to the child's skills but affect the production of the child's skills. The vector of investments I_t, in contrast, includes only those factors that are endogenous to the child's skills. Parents (or teachers or other subjects in the child's environment or the child herself) might invest more in the case the child's skill level is already high than in the case the skill level is low (reinforcement behavior). Or, conversely, investments might be higher in the case the child's skill level is low than in the case the child's skill level is already high (compensational behavior). To explicitely model the endogeneity of investments, I formulate the level of investments as a function of the child's skills as well as X:

$$I_t = f_2(\theta_t, X) \tag{8}$$

The positive production effect that I estimated in the empirical part of this paper could thus either be the result of WM capacity positively affecting the level of investments I_t in equation (8) and I_t positively affecting skill development in equation (7). Alternatively, even if WM capacity did not affect the level of investments, the positive production effect could be due to the cross derivative of equation (7) being positive, i.e., $\frac{\partial^2 \theta_{t+1}}{\partial \theta_t \partial I_t} > 0$. The latter is what Cunha and Heckman (2007) called 'dynamic complementarity'.

In my main analysis I do not model (nor measure) investments and therefore do not differentiate between the two mechanisms. I instead estimate the total effect. This is by purpose and due to the following reasoning:

Any activity of a child, being read a book, playing at the playground, talking to her parents, even sleeping, can affect her skills. This means that every minute in a child's life is an investment into skills, with positive or negative returns. Classifying a child's activities into two categories, one being investments into skills and the other being no investments into skills will always be arbitrary. But given that every minute is an investment, it is impossible to *increase* the quantity of investments, one can only *replace* an investment by another investment. But if on cannot increase the quantity of investments but only their quality, it becomes close to impossible (not only empirically but even conceptually) to differentiate between the two mechanisms, i.e., to answer the question of whether a skill shock increased the quality of an investment and as a consequence improved further skill production, or whether it made the investment (at the given quality) more productive.

Consider the following examples: First, take a child that—due to some exogenous skill improvement—becomes faster in finishing her homework and as a consequence spends more

time on reading her favorit book and therefore becomes better in reading. Nobody has changed the amount of time spent with the child, nor changed the material spendings for the child. The amount of time the child spends on her own and the materials available to the child have remained the same, but she spends her time now in a different way. Is this now an improvement of the quality of the investments (in terms of the child's own investment: she reads instead of spending long time on homework) or did the productivity of a given investment increase (the parents continue letting her spend one hour on her own but the time has become more productive)?

Take as a second example a child that—due to improved skills—has improved in chess and thus spends more of her freetime in playing chess with her friend (instead of, say, playing cards). Playing chess might improve her reasoning and concentration abilities. What is the mechanism for the initial skill improvement raising reasoning and concentration abilities? Is it a change of investments (playing chess instead of cards) or has just the productivity of the given investment (parents let the child play with her friend) improved?

Take as a third example a child that likes playing with Lego bricks together with her parent. After an exogenous skill enhancement the child and her parent construct more sophisticated Lego buildings than they would in absence of the treatment. This construction experience makes the child improve in visual-spatial imagination. Is this now a change of investments or are given investments (given time the parent spends playing Lego with her child) more productive?

The examples illustrate that in many cases, differentiating between the two mechanisms will be artificial and I therefore refrain from doing so. The core question in the context of the skill production function is whether improved skills can change the subsequent growth path of skills or not. This is the question I answer. And I provide evidence that this is the case. I conclude that the skill formation process is dynamic and that improving one skill not only raises the future level of this same skill but can even affect the subsequent production of other skills over time.

6 Conclusion

In this paper I have formulated a dynamic model of skill formation accounting for the multiplicity of skills. Based on the experimentally manipulated level of one skills—working memory (WM) capacity—I have estimated self- and cross-productivity effects. I found WM capacity to be self-productive, but only to the extent that a level shift persists over time without improving the production of further WM capacity. Furthermore, I found WM capacity to be cross-productive, i.e., improving the production of other skills (geometry skills, inhibition control, Raven's IQ). Accounting for the multiplicity of skills results to be important as I have documented strong cross-production effects on some skills (geometry, Raven's IQ, inhibition control), while none on others (arithmetic skills, reading skills, sustained concentration).

My finding of positive self- and cross-productivity contributes to the understanding of the skill formation process and provides an explanation for skill gaps widening over the life cycle.

The existence of self- and cross-productivity effects implies that educational interventions can have increasing effects on human capital over time. Given this increasing pattern, policy measures intended to foster human capital are likely to be more effective, the earlier in the life cycle they intervene.

Although I found positive production effects of WM capacity over a span of one year, I certainly do not claim this effect to be stable until infinity or at least through a very long period of time. But even if the identified cross-productivity effect existed only this one year, the improved cross-skills could have long-term implications through their own impact on the dynamic skill production process. Changes in the children's environment could also be a consequence of the improved cross-skills. The important lesson from my findings is that dynamic effects of educational interventions play a role and thus the timing and interdependency of skills are fundamental as well as the choice of the target skill of any intervention. The dynamic processes might explain why in our main analysis of the treatment effect (Berger et al. 2020) we found an impact of the 5-week intervention even on the longer term school career of children. The contribution of this paper is thus to shed more light on the mechanism through which the intervention operates in the longer term through dynamic self- and cross-productivity.

Certainly, the extent to which skills are self- and cross-productive most likely varies across skills, i.e., on the type of skill improved in the initial stage (Bailey et al. 2020) and the effects cannot be generalized. I speculate that more basic skills, such as working memory or basic reasoning skills, have stronger cross-productive effects on other skills than applied skills, such as calculation skills or drawing skills. The WM intervention study focussing on WM capacity, a basic skill needed for many tasks and activities (Baddeley 1999), thus is a suitable opportunity for studying self- and cross-productivity. In a similar vein, Bailey et al. (2017) argue that interventions should target what they call "trifecta" skills—ones that are malleable, fundamental, and would not have developed in the absence of the intervention. In order to be able to design effective and efficient educational interventions, one need to know what skills can be improved most effectively at the first stage by interventions and what types of skills improve the growth path of other skills strongest. Hence, further research about the most productive skills is needed.

References

Aksayli, N. D., G. Sala, and F. Gobet (2019). The cognitive and academic benefits of cogmed: A meta-analysis. *Educational Research Review 27*, 229–243.

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

Almlund, M., A. L. Duckworth, J. Heckman, and T. Kautz (2011). Personality psychology and economics. In *Handbook of the Economics of Education*, Volume 4, pp. 1–181. Elsevier.

Almond, D., J. Currie, and V. Duque (2018). Childhood circumstances and adult outcomes: Act ii. *Journal of Economic Literature 56*(4), 1360–1446.

Andersen, S. C. and H. S. Nielsen (2016). Reading intervention with a growth mindset approach improves children's skills. *Proceedings of the national academy of sciences 113*(43), 12111–12113.

Backes-Gellner, U., H. Herz, M. Kosfeld, and Y. Oswald (2018). Do preferences and biases predict life outcomes? Evidence from education and labor market entry decisions. CEPR Discussion Paper No. DP12609.

Baddeley, A. (1999). *Essentials of Human Memory*. Cognitive psychology. Psychology Press.

Bailey, D., G. J. Duncan, C. L. Odgers, and W. Yu (2017). Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of Research on Educational Effectiveness 10*(1), 7–39.

Bailey, D. H., G. J. Duncan, F. Cunha, B. R. Foorman, and D. S. Yeager (2020). Persistence and fade-out of educational-intervention effects: Mechanisms and potential solutions. *Psychological Science in the Public Interest 21*(2), 55–97.

Berger, E. M., E. Fehr, H. Hermes, D. Schunk, and K. Winkel (2020). The impact of working memory training on children's cognitive and noncognitive skills. Discussion Paper 13338, IZA Institute of Labor Economics.

Bergman Nutley, S. and S. Söderqvist (2017). How is working memory training likely to influence academic performance? current evidence and methodological considerations. *Frontiers in psychology 8*.

Blomeyer, D., M. Laucht, K. Coneus, and F. Pfeiffer (2009). Initial risk matrix, home resources, ability development, and children's achievement. *Journal of the European Economic Association 7*(2-3), 638–648.

Borghans, L., A. L. Duckworth, J. J. Heckman, and B. Ter Weel (2008). The economics and psychology of personality traits. *Journal of human Resources 43*(4), 972–1059.

Bowles, S., H. Gintis, and M. Osborne (2001). The determinants of earnings: A behavioral approach. *Journal of Economic Literature 39*(4), 1137–1176.

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

Carrell, S. E., M. Hoekstra, and E. Kuka (2018). The long-run effects of disruptive peers. *American Economic Review 108*(11), 3377–3415.

Chiteji, N. (2010). Time preference, noncognitive skills and well being across the life course: Do noncognitive skills encourage healthy behavior? *American Economic Review: Papers & Proceedings 100*, 200–204.

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

Cunha, F. and J. J. Heckman (2008). Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. *Journal of Human Resources 43*(4), 738–782.

Cunha, F., J. J. Heckman, L. J. Lochner, and D. V. Masterov (2006). Interpreting the evidence on life cycle skill formation. In E. A. Hanushek and F. Welch (Eds.), *Handbook of the Economics of Education*, pp. 697–812. Amsterdam, North-Holland: Elsevier.

Cunha, F., J. J. Heckman, and S. M. Schennach (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica 78*(3), 883–931.

Currie, J. and D. Almond (2011). Human capital development before age five. In *Handbook of labor economics*, Volume 4, pp. 1315–1486. Elsevier.

Dohmen, T., A. Falk, D. Huffman, and U. Sunde (2009). Homo reciprocans: Survey evidence on behavioural outcomes. *The Economic Journal 119*(536), 592–612.

Eigsti, I.-M., V. Zayas, W. Mischel, Y. Shoda, O. Ayduk, M. B. Dadlani, M. C. Davidson, J. L. Aber, and B. Casey (2006). Predicting cognitive control from preschool to late adolescence and young adulthood. *Psychological Science 17*(6), 478–484.

Elsner, B. and I. E. Isphording (2017). A big fish in a small pond: Ability rank and human capital investment. *Journal of Labor Economics 35*(3), 787–828.

Esser, G., A. Wyschkon, and K. Ballaschk (2008). *Basisdiagnostik Umschriebener Entwicklungsstörungen im Grundschulalter (BUEGA)*. Göttingen: Hogrefe.

Falk, A., F. Kosse, P. R. Dovern-Pinger, H. Schildberg-Hörisch, and T. Deckers (2020). Socio-economic status and inequalities in children's iq and economic preferences. *Journal of Political Economy* (forthcoming).

Fiorini, M. and M. P. Keane (2014). How the allocation of children's time affects cognitive and noncognitive development. *Journal of Labor Economics 32*(4), 787–836.

Frison, L. and S. J. Pocock (1992). Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design. *Statistics in Medicine 11*(13), 1685–1704.

Gawrilow, C. and P. M. Gollwitzer (2008). Implementation intentions facilitate response inhibition in children with adhd. *Cognitive Therapy and Research 32*(2), 261–280.

Gibbons, S., O. Silva, and F. Weinhardt (2017). Neighbourhood turnover and teenage attainment. *Journal of the European Economic Association 15*(4), 746–783.

Hanushek, E. A., G. Schwerdt, S. Wiederhold, and L. Woessmann (2015). Returns to skills around the world: Evidence from piaac. *European Economic Review 73*, 103–130.

Hart, B. and T. R. Risley (1995). *Meaningful differences in the everyday experience of young American children.* Paul H Brookes Publishing.

Heckman, J. and P. Carneiro (2003). Human capital policy. In J. J. Heckman, A. B. Krueger, and B. Friedman (Eds.), *Inequality in America: What role for human capital policies?*, pp. 77–239. The MIT Press.

Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science 312*(5782), 1900–1902.

Heckman, J. J. (2007). The economics, technology, and neuroscience of human capability formation. *Proceedings of the National Academy of Sciences 104*(33), 13250–13255.

Heckman, J. J., J. E. Humphries, and G. Veramendi (2018a). The nonmarket benefits of education and ability. *Journal of human capital 12*(2), 282–304.

Heckman, J. J., J. E. Humphries, and G. Veramendi (2018b). Returns to education: The causal effects of education on earnings, health, and smoking. *Journal of Political Economy 126*(S1), S197–S246.

Heckman, J. J. and T. Kautz (2014). Fostering and measuring skills: Interventions that improve character and cognition. In J. J. Heckman, J. Humphries, and T. Kautz (Eds.), *The Myth of Achiement Tests: The GED and the Role of Character in American Life*, pp. 341–430. University of Chicago Press.

Heckman, J. J., J. Stixrud, and S. Urzua (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics 24*(3), 411–482.

Helmers, K. F., S. N. Young, and R. O. Pihl (1995). Assessment of measures of impulsivity in healthy male volunteers. *Personality and Individual Differences 19*(6), 927–935.

Holmes, J., S. E. Gathercole, and D. L. Dunning (2009). Adaptive training leads to sustained enhancement of poor working memory in children. *Developmental science 12*(4).

Jackson, C. K. (2018). What do test scores miss? The importance of teacher effects on non–test score outcomes. *Journal of Political Economy 126*(5), 2072–2107.

Kane, T. J., E. S. Taylor, J. H. Tyler, and A. L. Wooten (2011). Identifying effective classroom practices using student achievement data. *Journal of human Resources 46*(3), 587–613.

Kautz, T., J. J. Heckman, R. Diris, B. Ter Weel, and L. Borghans (2014). Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success. Technical report, National Bureau of Economic Research.

Knudsen, E. I., J. J. Heckman, J. L. Cameron, and J. P. Shonkoff (2006). Economic, neurobiological, and behavioral perspectives on building America's future workforce. *Proceedings of the National Academy of Sciences 103*(27), 10155–10162.

Kosse, F., T. Deckers, P. Pinger, H. Schildberg-Hörisch, and A. Falk (2020). The formation of prosociality: causal evidence on the role of social environment. *Journal of Political Economy 128*(2), 434–467.

Lindqvist, E. and R. Vestman (2011). The labor market returns to cognitive and noncognitive ability: Evidence from the swedish enlistment. *American Economic Journal: Applied Economics 3*, 101–128.

Martinussen, R., J. Hayden, S. Hogg-Johnson, and R. Tannock (2005). A meta-analysis of working memory impairments in children with attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child & Adolescent Psychiatry 44*(4), 377–384.

McKenzie, D. (2012). Beyond baseline and follow-up: The case for more t in experiments. *Journal of Development Economics 99*(2), 210–221.

Melby-Lervåg, M. and C. Hulme (2013). Is working memory training effective? a meta-analytic review. *Developmental psychology 49*(2), 270.

Melby-Lervåg, M., T. S. Redick, and C. 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 on Psychological Science 11*(4), 512–534.

Moffitt, T. E., L. Arseneault, D. Belsky, N. Dickson, R. J. Hancox, H. Harrington, R. Houts, R. Poulton, B. W. Roberts, S. Ross, et al. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences 108*(7), 2693–2698.

Murphy, R. and F. Weinhardt (2020). Top of the class: The importance of ordinal rank. *Review of Economic Studies forthcoming*.

Sala, G. and F. Gobet (2020). Working memory training in typically developing children: A multilevel meta-analysis. *Psychonomic Bulletin & Review*, 1–12.

Shipstead, Z., K. L. Hicks, and R. W. Engle (2012). Cogmed working memory training: Does the evidence support the claims? *Journal of Applied Research in Memory and Cognition 1*(3), 185–193.

Sorrenti, G., U. Zölitz, D. Ribeaud, and M. Eisner (2020). The causal impact of socio-emotional skills training on educational success. IZA Discussion Paper 13087.

Online Appendix

A Supplementary Details on the Data

Since I use the same data source for the empirical part of this paper as I did with my co-autrhos in Berger et al. (2020), the study description in this section is—in large parts—a reproduction of the relevant parts of the appendix to Berger et al. (2020). This is done for the reader's convenience.

The experimental 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 A.1), the treatment conditions (Section A.2), the data collection waves (Section A.3), and the outcome measures (Section A.4). 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).

A.1 Participants

A.1.1 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 in the treatment classes, which we would randomly choose in the following step and (ii) the children (from both treatment and control classes) 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.

A.1.2 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 each. 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%.[5] We were able to evaluate 572 children of the 580 for whom we received parental consent to collect data for our final data set.[6] 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.

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-Lervåg and Hulme (2013)), the attrition rate was 10–11% even though the last follow-up measurements took place only 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. The analysis sample used in this paper is restricted to the sample including nonmissing data both of test outcomes measured in W4 (because these are the main outcome measures for this paper) and of the WM score measured in W2 (because this is the main explanatory variable in this paper), it thus contains 518 observations. Among this sample, 261 children were girls (50.4%) and 257 were boys (49.6%). Mean age at the beginning of the year (January 1, 2013) (standard deviation = 0.36 years).

A.2 The Treatment

A.2.1 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 WM training. 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

[5] Among the children for whom we did not receive parental consent, roughly 50% participated in the WM training while the other roughly 50% were in the control classes. Due to the lacking parental consent we could not collect data in W1–W4 for these children. Despite the lacking parental consent for the data collection, the children did participate in the WM training because the school authorities viewed the training as part of regular teaching.

[6] Among the 572 children, six children (two of them in the control group) completed the baseline (W1) tests slightly after the actual start of the WM training because they were sick or absent at the regular test date. Since the delays were rather small, we kept these children in the sample. All reported effects of WM training remain intact if we exclude these children from the data analysis.

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.[7] The assistants also helped in preparing a comprehensive documentation of the training.

In the same sample of children other treatments (unrelated to WM training) were conducted with a randomly chosen part of the WM treatment group and a randomly chosen part of the control group. Since the other treatments were orthogonal to the WM treatment, this should not affect our results. Nevertheless, we carried out a robustness test controlling for the other treatments. Our results are unchanged compared to our baseline results.

A.2.2 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 during the time of the intervention. The control group had no access to the WM training software.

Children in the treatment group worked with the external mouse during the training. This ensured 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 A.3).

A.2.3 Software

The WM training software used for the treatment was "Cogmed RM"[8] in an offline version with German instructions. It provides an age-specific user-interface, adaptive levels of difficulty, and a

[7] The assistants were university students who were familiar with the WM training software.

[8] Cogmed and Cogmed Working Memory Training are trademarks, in the U.S. and/or other countries, of Cogmed Inc. (www.cogmed.com).

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. This is consistent to the actual findings of Berger et al. (2020).

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 within the training program 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).[9] 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 thus no further training was 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).

A.3 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

[9]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 into account the time needed for picking up and bringing back the notebooks).

(W4). In each wave, we conducted several computer-based tests that served the purpose of measuring the consequences of WM training on skills. We describe these tests in detail below. In addition, we administered questionnaires to teachers and parents. In W4, we also asked the children a few questions after they had finished 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.

A.3.1 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.

A.3.2 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%).

A.3.3 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. Children's mean age on the day the teacher questionnaire was submitted equals to 85.4 months in W1, 88.1 months in W2, 92.9 months in W3, and 100.3 months in W4. We achieved a 100% return rate for the teacher questionnaire in all four evaluation waves.

A.4 Outcome Measures

This section describes the test measures that we use in our analysis. The main study (Berger et al. 2020) contains three WM tests. Since only one of these measures visuo-spatial WM capacity, we focus on that one. 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 (Bulheller and Häcker 2010) 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 et al. 2008)). For the ease of interpretation and comparison, we standardize all test scores to mean = 0 and standard deviation = 1, separately by test and wave and based on the control group. Histograms of the distribution of all raw test scores (i.e., before standardization) for the evaluation waves W1–W4 are displayed in Figures A1–A4.

A.4.1 Working Memory Test

WM capacity was measured by a visuo-spatial complex span task. To avoid task-learning effects, we chose a task distinct from the training tasks. In the task, 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 A5). Then, a new stimulus screen appeared and the child again had to click on the deviant shape, etc. Figure A5 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 A5, for example, the

Figure A1: Distribution of Nonstandardized W1 Test Scores

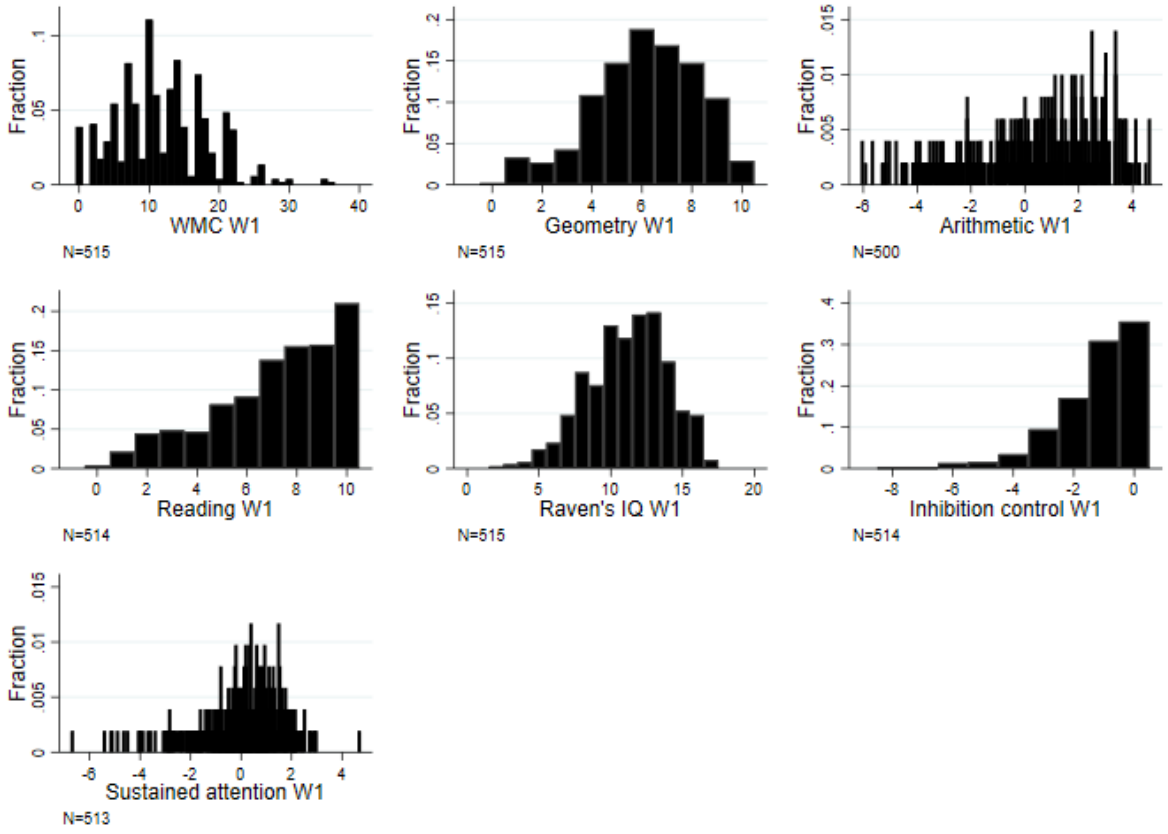


Figure A2: Distribution of Nonstandardized W2 Test Scores

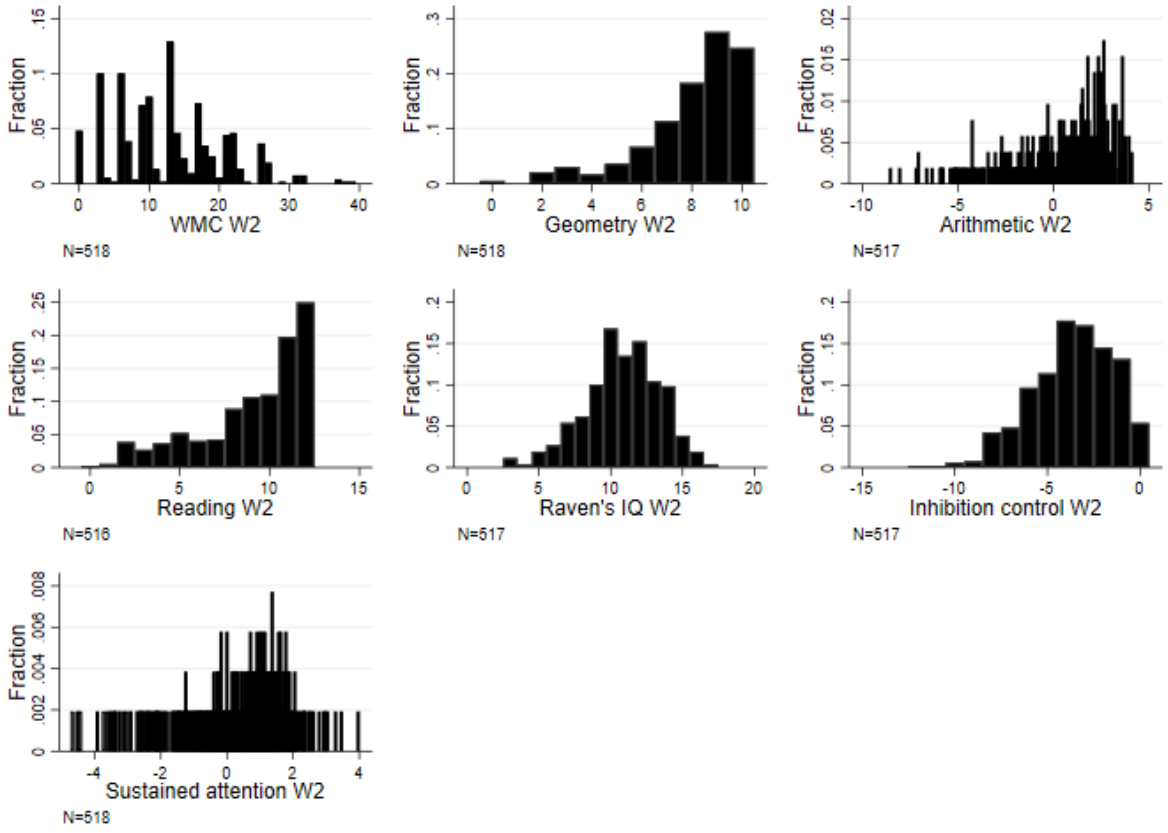


Figure A3: Distribution of Nonstandardized W3 Test Scores

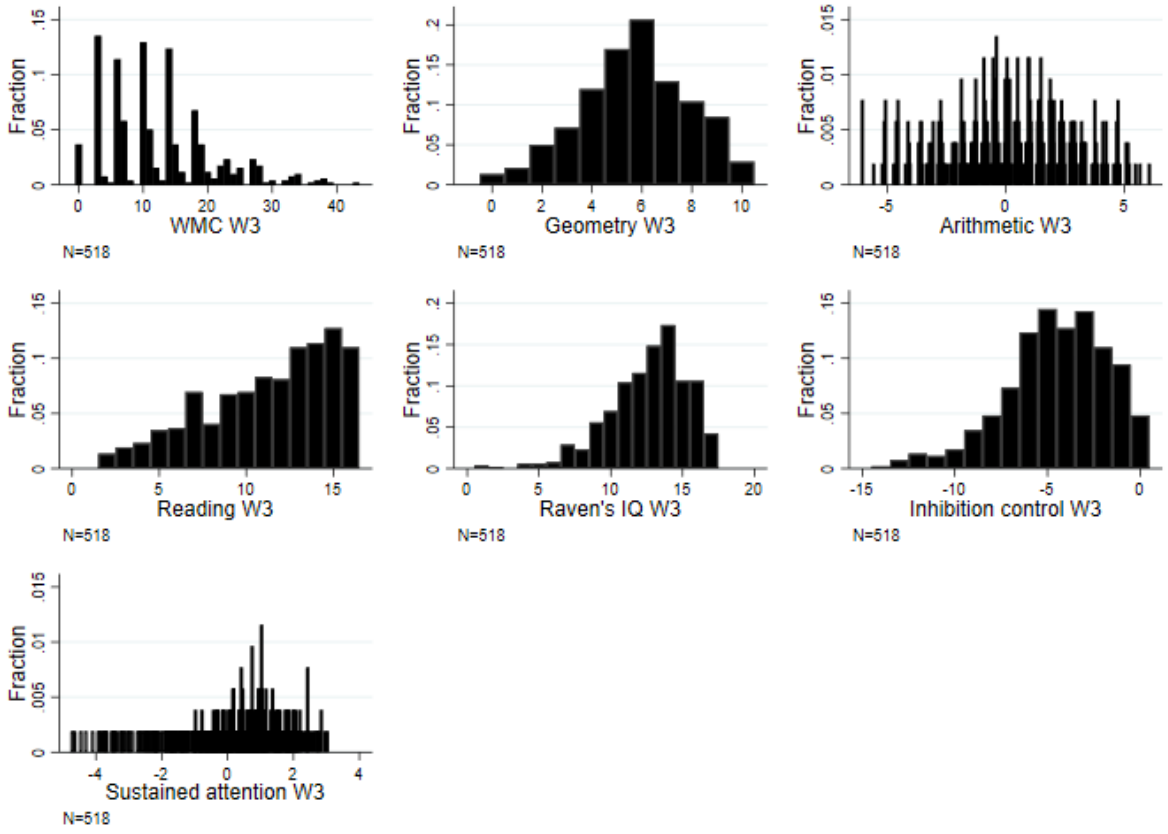
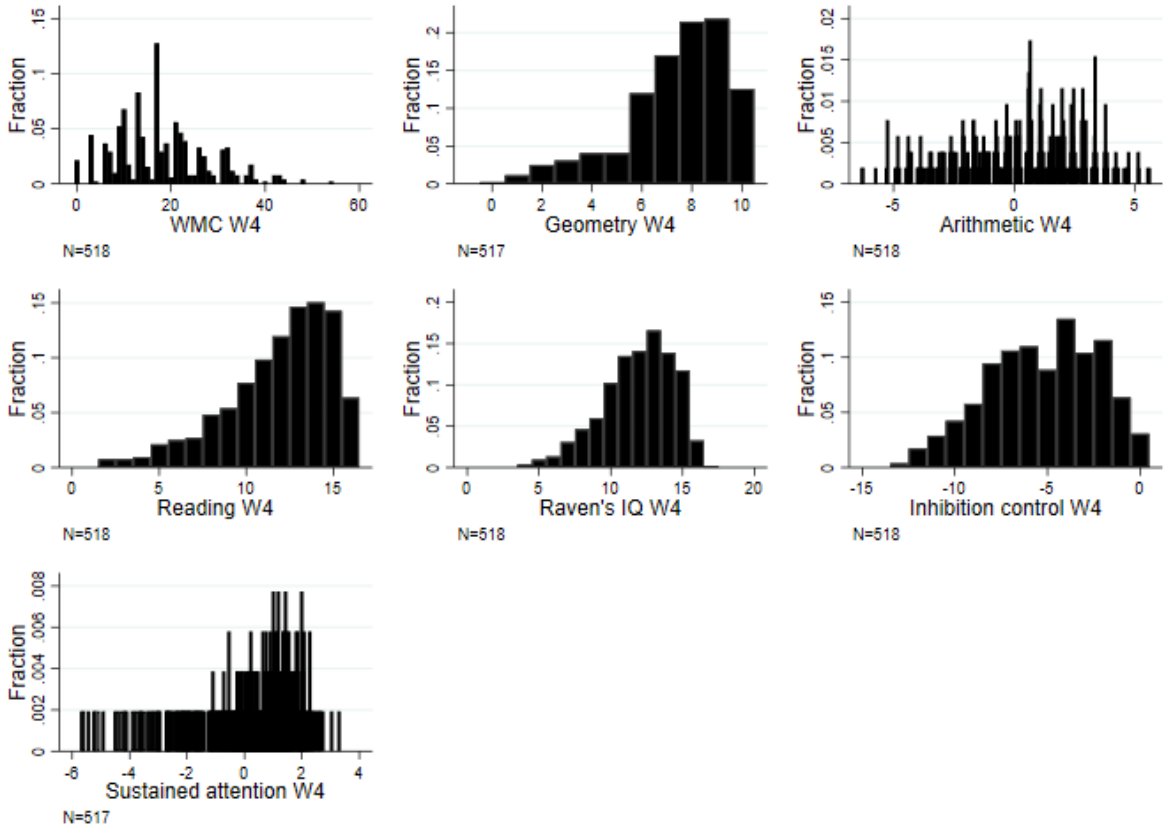
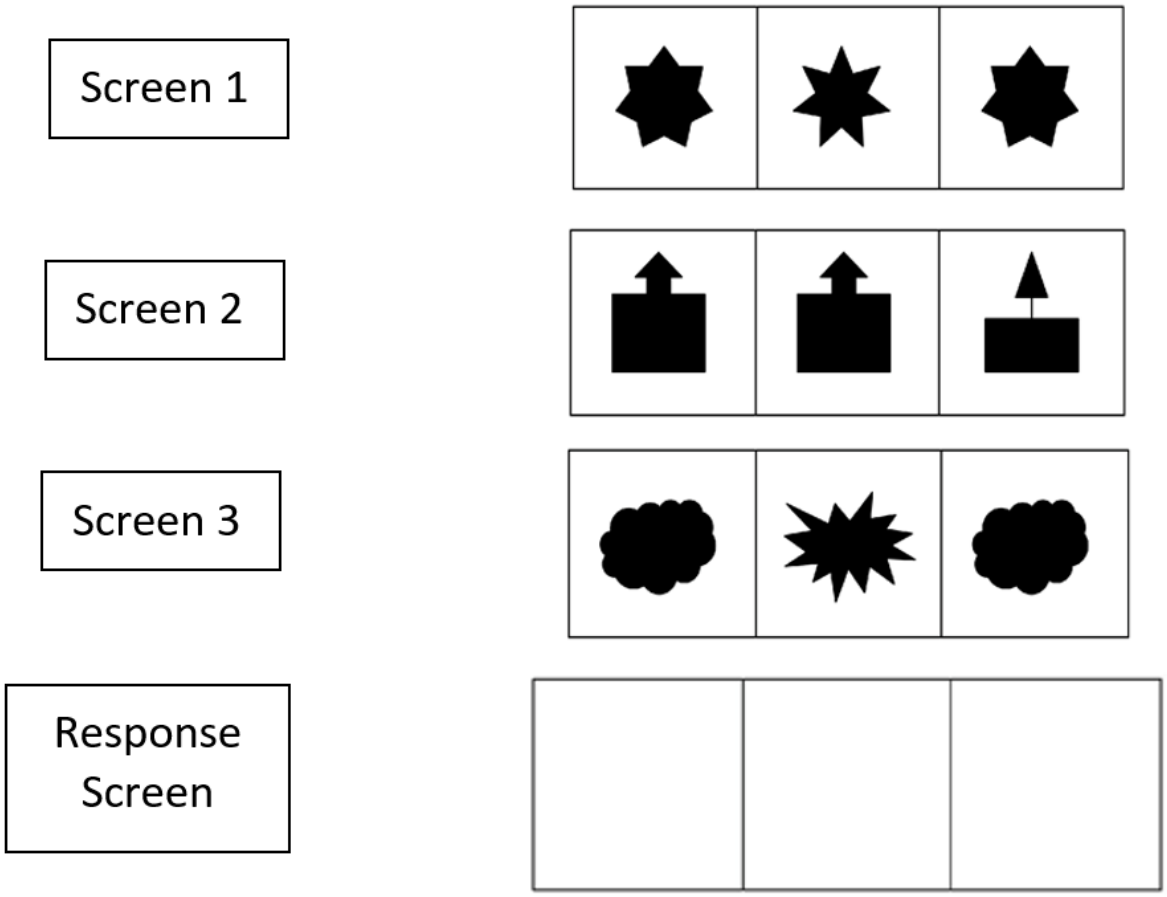


Figure A4: Distribution of Nonstandardized W4 Test Scores



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 A5: Stimulus and Response Screens in the 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). We standardized this score to mean = 0 and standard deviation = 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.

A.4.2 Educational Achievement Tests

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 A6). 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 A6: 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 A7 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.

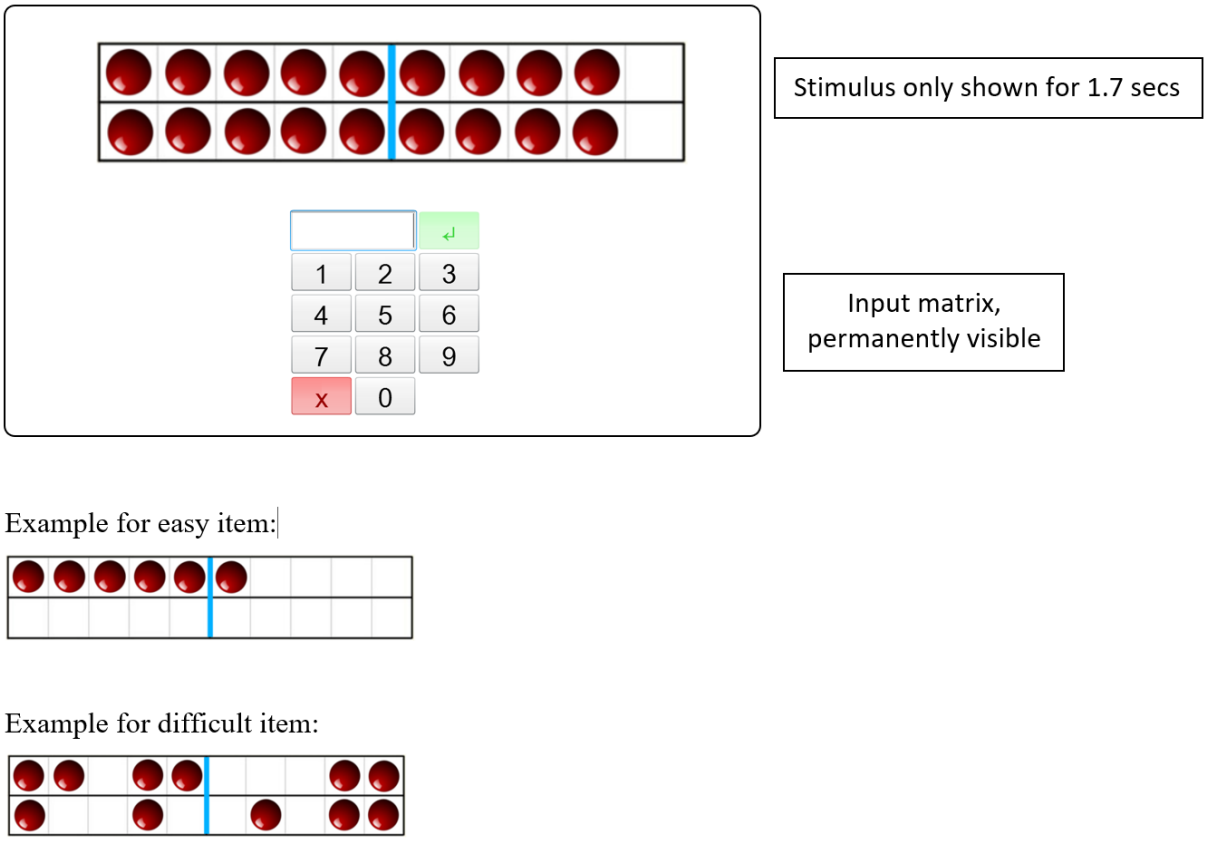
A two by ten grid with the subdivision at 5 is used in the first grade in the participating primary schools to teach numbers and calculations. To solve the number sense task, children need to be familiar with the number range up to 20, and a good understanding of the logic of the grid is useful. Since 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 A7) 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 A7) 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.

Figure A7: Number Sense Task, Screenshot Plus Two Further Examples

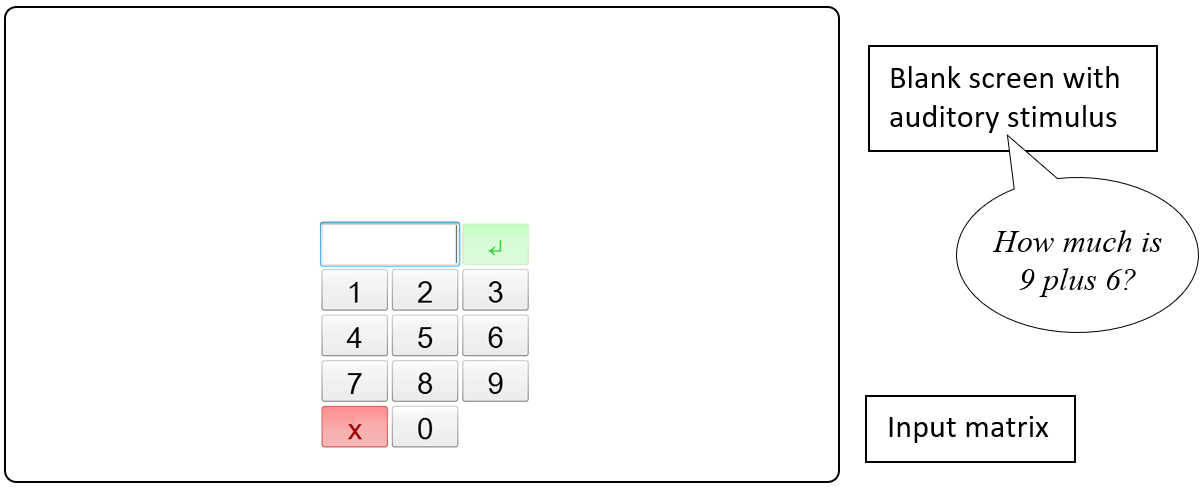


Auditory arithmetic task

This subtask measures arithmetic skills for addition and subtraction of two numbers (see Figure A8). Computational tasks were presented over the headphone (e.g., "How much is 9 plus 6?"). Children had to listen and 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.

Figure A8: Auditory Arithmetic Task, Screenshot Plus Two Further Examples



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

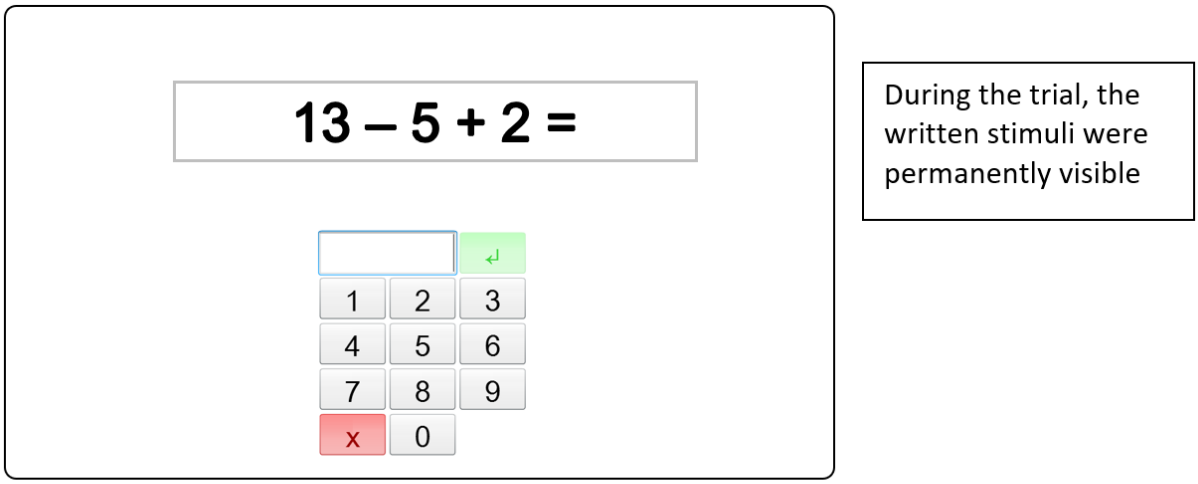
Example for difficult item:
"How much is 92 minus 17?"

Written arithmetic task

In contrast to the auditory task, the arithmetic problems in the written 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 A9). 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.

Figure A9: Written Arithmetic Task, Screenshot Plus Two Further Examples



Example for easy item:

1 + 5 + 4 =

Example for difficult item:

100 – 43 – 20 + 43 =

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

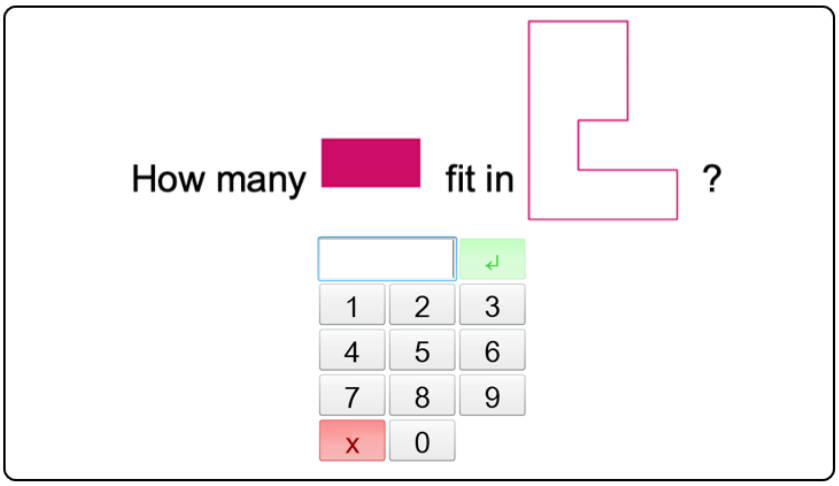
Computation of final arithmetic test score

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 standard deviation = 1 within each wave. We then added up the three standardized subscores and standardized this composite score to mean = 0 and standard deviation = 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 A10 below). Depending on the size and the shape of the larger geometric object, this task can be made harder or easier.

Figure A10: Geometry Task, Screenshot Plus Two Further Examples



Stimulus and
Input matrix
permanently visible

Example for easy item:

Example for difficult item:

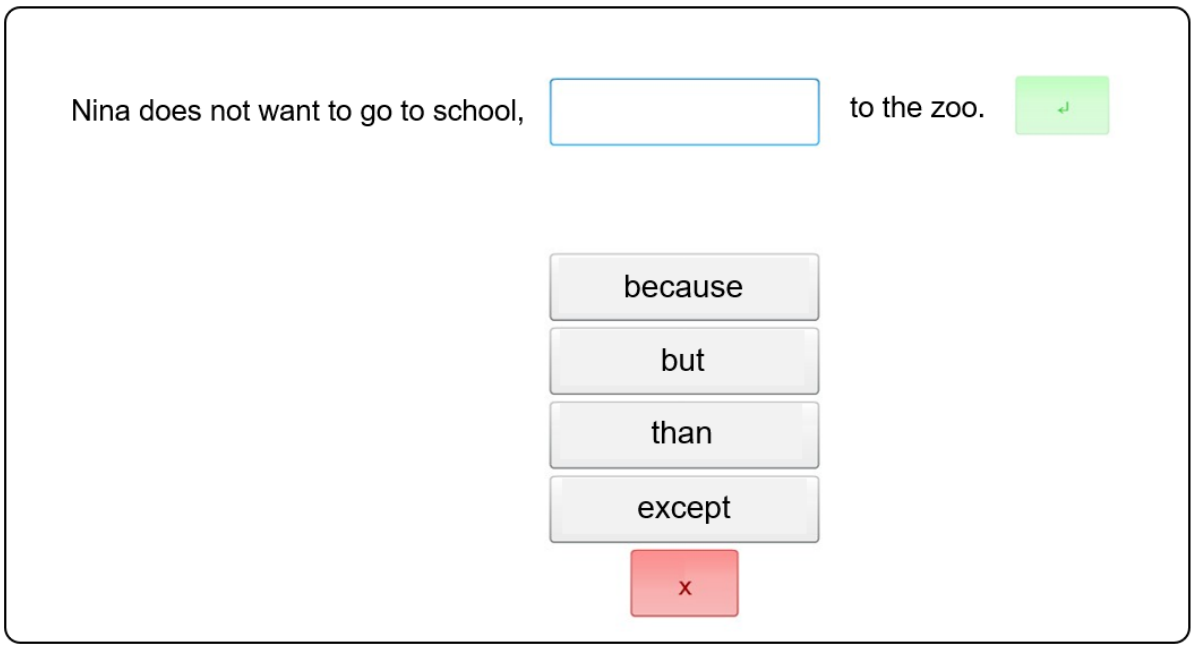
The task contained 10 items in each evaluation wave. The difficulty level varied across items and evaluation waves. Difficulty varied along various dimensions. Consider the easy item shown in Figure A10 (the red square): Children could solve the problem without any mental rotation of the small square. Furthermore, the larger object is subdivided into two components, making the task even easier. In contrast, for the first item shown in Figure A10 (the pink rectangle), children had to mentally rotate the small object to solve the question. For the difficult item in Figure A10 (the green triangle), children hat to mentally rotate the triangle, store the number for subparts and keep track of which parts were already counted when filling the larger geometric object.

We constructed an outcome score by counting the correctly solved items and standardizing the figure to mean = 0 and standard deviation = 1 within each wave.

Reading comprehension skills

Reading comprehension was assessed by a sentence comprehension test in single choice format. On the screen (see Figure A11), 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 tipping on the green enter button right beside the sentence.

Figure A11: 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 sentences containing 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.

We again constructed the outcome score by counting the correctly solved items and standardizing the figure to mean = 0 and standard deviation = 1 within each wave.

A.4.3 Fluid IQ

Children's fluid IQ was measured using a set of Colored Progressive Raven's Matrices (Bulheller and Häcker 2010). 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.

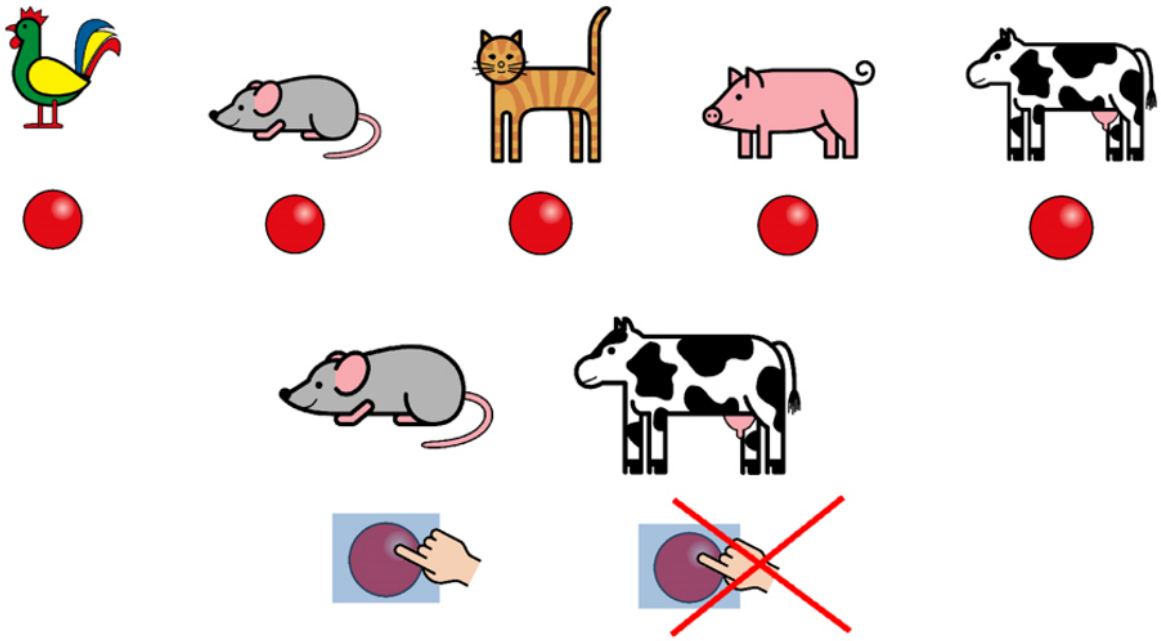
A.4.4 The Go/No-Go Task

To measure inhibitory abilities, we employed a go/no-go task that was adapted from 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 A12 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 (the display 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 red button. 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 standard deviation = 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.

Figure A12: The Animals and the "Go-Button" in the Go/No-Go Task

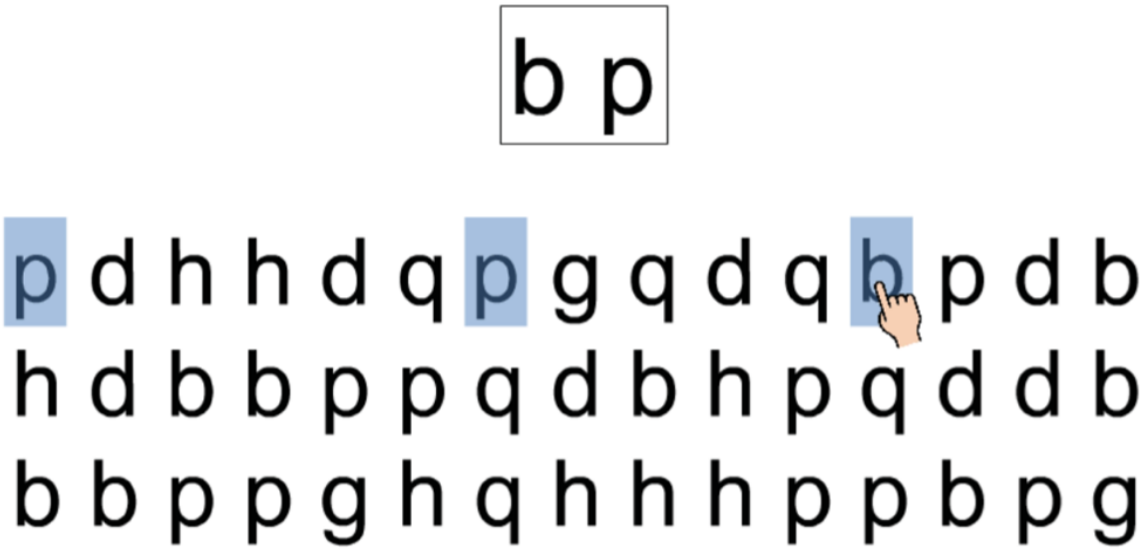


A.4.5 The bp Task

The bp task measures sustained concentration and is taken from Esser et al. (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 A13 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 standard deviation = 1 within each wave and multiplied by -1. Thus, a higher score indicates better performance in the task (i.e., fewer mistakes).

Figure A13: Example of a Screen in the bp Task



B Supplementary Robustness Tests

Table A1: 2SLS estimate of cross-productivity on the d'-score based on the go/no-go task

	d'-score
WMC W2	0.465
	(0.324)
N	514

The result is based on a two stage least square model using the d'-score measured in W4 as dependent variable and the treatment indicator as instrument for working memory capacity (WMC) in W2. The model further includes school fixed effects as well as controls for gender, age, and the pre-treatment d'-score. The skill variables are standardized to mean = 0 and standard deviation = 1. Standard errors in parentheses are clustered on the classroom level. * p<0.05.

Table A2: 2SLS estimates of self- and cross productivity—dependent variables measured in W3

	(1) WMC	(2) Geometry	(3) Arithmetic	(4) Reading	(5) Raven's IQ	(6) Inhibition control	(7) Sustained attention
WMC W2	1.414*	0.165	0.092	0.043	0.870*	-0.093	0.408
	(0.287)	(0.265)	(0.284)	(0.328)	(0.170)	(0.179)	(0.225)
N	515	515	500	514	515	514	513

The results are based on two stage least square models using skill levels measured in W3 as dependent variables and the treatment indicator as instrument for WMC W2. The models further include school fixed effects, gender and age controls, and the pre-treatment skill score. The skill scores are standardized to mean = 0 and SD = 1. Standard errors in parentheses are clustered on the classroom level. * p<0.05, ** p<0.01.