Title	Working memory and numeracy training for children with math learning difficulties: Evidence from a large-scale implementation in the classroom
Author(s)	David Muñez, Kerry Lee, Rebecca Bull, Kiat Hui Khng, Fiona Cheam and Ridzuan Abd Rahim

Working memory and numeracy training for children with math learning difficulties: evidence from a large-scale implementation in the classroom

David Muñez[1], Kerry Lee[2], Rebecca Bull[3], Kiat Hui Khng[1], Fiona Cheam[4] and Ridzuam Abd Rahim[4]

[1]Centre for Research in Child Development, National Institute of Education, Nanyang Technological University, Singapore

[2]Dept. of Early Childhood Education, The University of Education Hong Kong, Hong Kong

[3]Dept. of Educational Studies, Macquarie University, Australia

[4]Ministry of Education, Singapore

Acknowledgments: This study was funded by Singapore Ministry of Education (MOE) under the Education Research Funding Programme and administered by National Institute of Education, Nanyang Technological University, Singapore. The views expressed in this paper are the author's and do not necessarily represent the views of NIE.

Abstract

We explored the challenges, limitations, and potential effectiveness of a large-scale computerized working memory and numeracy intervention in the classroom with children at risk of mathematical learning disabilities (n= 428, M_{age}= 83.85 months, 41% female). Children were assigned to four different treatment protocols (working memory [WM], working memory plus numeracy [NWM], numeracy [NUM], and active control [AC]) that were implemented as part of normally scheduled class activities for one year. Wide variability in training exposure highlighted the challenges of implementing an ecologically valid large-scale classroom intervention. The NUM and NWM intervention contributed to improvements in various early numeracy skills as well as math achievement after accounting for training exposure. Some of these effects emerged once the intervention concluded. However, the intervention failed to improve WM, which was likely due to insufficient training dosage in the practical setting. Findings suggest that combining both working memory and numerical skills training is worth further investigation. The study also provides evidence of challenges related to the implementation of training programs in real-life learning environments.

Keywords: mathematical disabilities, working memory training, numeracy training, intervention

Educational Impact And Implications Statement

Studies examining working memory (WM) and numeracy training programs for children at risk of math learning difficulties have increased exponentially during the last decade. The current findings suggest that numeracy training programs may benefit these children to a larger extent. Findings also call attention to the translation of lab-based training programs into tools to support teaching efforts and how to overcome the inherent limitations of real classroom settings.

Working memory and numeracy training for children with math learning difficulties: evidence from a large-scale implementation in the classroom

Epidemiological studies suggest that approximately 6%–14% of school-age children have persistent difficulty with mathematics despite adequate learning opportunities and age-appropriate achievement in other domains (Von Aster & Shalev, 2007; prevalence studies suggest that about 5% - 6% of children may show some kind of math learning difficulty, Devine et al., 2013). Heightened interest in the nature and origins of these learning difficulties has resulted in increased research on how to provide adequate support to children with mathematics learning difficulties. The main objective of the current research is to determine the effectiveness of a working memory (WM) and numeracy training intervention for children with mathematics learning difficulties as well as illustrating the challenges and limitations of large-scale interventions in real-life learning environments. This is particularly relevant because the bulk of evidence supporting WM and numeracy interventions comes from lab-settings or demands administration protocols that do not align with the dynamics and settings of real-life learning environments.

Mathematics learning disabilities—a blended disorder

The term math learning difficulties (MLD) is broadly used to define difficulties with a variety of math skills that cannot be explained by low intelligence, neurological disorders, or inadequate instruction. Despite an increasing body of research on MLD, its causes remain unknown and the patterns of difficulties that characterize children and adults with MLD are still discussed (for reviews, see Bartelet, Ansari, Vaessen & Blomert, 2014; Nelson & Powell, 2017). Among the studies that have tried to parcel out the types and origins of MLD, Karagiannakis, Baccaglini-Frank, and Papadatos (2014) have suggested four groups of cognitive impairments: core number knowledge, memory retrieval and processing, reasoning, and visuospatial difficulties. Thus, difficulties with math are linked to specific cognitive

constructs and mechanisms, such as working memory (WM), long term memory (semantic memory), executive functions, fact-retrieval, and numerical magnitude processing, that are activated in the context of mathematical performance (Fuchs et al., 2005; Swanson et al., 2008).

Among these mechanisms, WM is one of the most often referred cognitive process underlying MLD (for reviews, see Peng & Fuchs, 2016; Szűcs, 2016). According to Baddeley (2012), working memory refers to the capacity to preserve while simultaneously processing the same or other information. The ability to update or refresh information efficiently is a key process in the executive component of WM and is thought to be a key ability that affects capacity. There is evidence that WM is a significant and unique correlate of latent classes of children with MLD (e.g., Swanson, Olide & Kong; 2018; see also, Swanson & Beebe-Frankenberger, 2004). Indeed, it has been suggested that children at risk of MLD can be identified at school entry on the basis of their poor WM capacity (Gathercole, Brown, & Pickering, 2003). WM is thought to be needed in the multiple steps needed to store, recall, and remember numerical information during math tasks.

Difficulties with numerical magnitude processing and the concept of number have also been found to characterize children with MLD (e.g., Fuchs et al., 2012; Geary. 2011). For instance, Bull, Lee and Muñez (2021; see also, Swanson et al., 2018) found that poor number line estimation skills in kindergarten characterized children at risk of developing MLD in elementary school. Indeed, MLD screeners feature tasks that evaluate basic skills like numerical comparison and non-symbolic-to-symbolic mapping (e.g., Butterworth & Laurillard, 2010; Nosworthy et al., 2013). These interrelations between cognitive and math difficulties have spurred research aimed at designing early interventions that focus on these particular numerical and cognitive aspects.

WM training interventions

With the well-established relations between WM and mathematics (e.g., Lee, Bull, & Ho, 2013; St. Clair-Thompson & Gathercole, 2006; Szűcs, Devine, Soltesz, Nobes, & Gabriel, 2013; for reviews, see Bull & Lee, 2014; Peng et al., 2016), WM or updating capacity has become a popular target for interventions. Though conceptually distinct, WM and updating are closely associated at the measurement level (Schmiedek et al., 2009; St Clair-Thompson & Gathercole, 2006; Wilhelm, Hildebrandt, & Oberauer, 2013). The idea that poor WM may place constraints on other higher cognitive functions suggests that, if training can enhance WM or updating (i.e., near-transfer effects), this should produce far-transfer effects to diverse untrained tasks that involve the activation of WM resources (Klingberg, Forssberg, & Westerberg, 2002; Klingberg et al., 2005).

Although most WM training programs have used adult participants (for a comprehensive review, see Melby-Lervåg & Hulme, 2013), there is evidence that WM training also triggers WM gains in children (Ramani et al., 2017; for a meta-analysis, see Scionti, Cavallero, Zogmaister, & Marzocchi, 2020). Indeed, training gains are expected to be more evident in children (Wass, Scerif, & Johnson, 2012), mainly because the rate of WM development strengthens significantly throughout childhood (Best & Miller, 2010; for a review, see Cowan, 2016). The bulk of evidence on WM training comes from Cogmed—a widely available commercial program that focuses on visuospatial WM training (Cogmed, 2005; for a recent review see, Aksayli, Sala, & Gobet, 2019). Nonetheless, there is also evidence from other types of WM trainings and paradigms. For instance, Ang, Lee, Cheam, Poon, and Koh (2015) found that updating training resulted in WM improvements in primary school students at immediate post-test (i.e., near-transfer effect), which were sustained and significant six months post-training.

Although different meta-analysis and systematic reviews suggest that WM training temporarily increases children's WM capacity (e.g., Sala & Gobet, 2017; Scionti et al., 2020), whether such improvements represent substantive increases in WM capacity remains controversial. Shipstead, Hicks, and Engle (2012), provided a comprehensive review of Cogmed studies and underscored several issues that included poor replicability of findings, methodological limitations such as the absence of control group, and crucially, conceptual issues that related to how WM improvements are frequently assessed in studies that have used Cogmed training. They found that most studies have used simple span tasks or short-term memory tasks that do not test the capacity of the executive component of WM, which has been found most closely related to math achievement. Indeed, this is an issue that affects the interpretation of findings from meta-analyses as WM improvements in aspects that have not been directly trained are sometimes called far-transfer effects, which is a term that is also used to define improvements in non-WM domains.

There is also a lack of consensus on whether WM improvements are stable over time. Indeed, recent reviews suggest that WM gains vanish once training is terminated (Kassai, Futo, Demotrovics, & Takacs, 2019; Scionti, et al., 2020; Takacs & Kassi, 2019; but see, Aksayli et al., 2019; Peijnenborgh et al., 2016; Schwaighofer et al., 2015, for opposite findings showing long-term effects for some trained skills).

Evidence is also mixed when it comes to how improvements in WM translate into improvements in math achievement (far-transfer effects). For instance, Söderqvist and Bergman-Nutley (2015), found that Cogmed training improved children's math achievement. Similarly, Holmes, Gathercole, and Dunning (2009), using Cogmed training with 10-year-olds, found an improvement in mathematical reasoning scores six months after training. However, because the control group in this study was not re-tested at a delayed post-test, it is unclear whether the improvement can be attributed to training (see also Kroesbergen, van't

Noordende, & Kolkman, 2014). Similar concerns arise from other studies that have not included a control group (St Clair-Thompson, Stevens, Hunt, & Bolder, 2010; Witt, 2011). Indeed, far transfer effects to academic outcomes have proven elusive under more rigorous experimental conditions such as RCT (random controlled trials). For instance, Dunning, Holmes, and Gathercole (2013), working with 8-year-olds with low WM, found Cogmed did not improve performance on standardized measures of reading and math (see also Roberts et al., 2016). In fact, the effects of WM training were not observed in classroom-based activities that resembled the training (e.g., following multi-step spoken instructions). Studies that have adhered to similar RCT designs but have considered typical populations have also shown null far-transfer effects (e.g., Hitchcock & Westwell, 2017). Other WM training programs have rendered similar results. Recently, in an attempt to replicate the far-transfer effect that was reported in Ramani, Jaeggi, Daubert, and Buschkuehl (2017), Ramani, et al., (2020) found that improvements in kindergarteners' WM capacity did not translate into improvements in their math skills (see also, Ang et al., 2015). Recent and more comprehensive meta-analyses and systematic reviews suggest that there is a weak link between WM improvements and math achievement in children (e.g., Kassai et al., 2019; Scionti et al., 2020).

Numeracy training interventions

The strength of the association between basic numeracy skills and later mathematics has also spurred efforts to support children who are struggling with math. Although some of these programs have focused on strengthening children's arithmetic skills (Fuchs et al., 2013; Gersten et al., 2015; van der Ven, Segers, Takashima, & Verhoeven, 2017)—since poor arithmetic abilities in formal school are one of the first signs of MLD, others have focussed on the building blocks of arithmetic (e.g., understanding the concepts of more and less, larger and smaller, matching symbolic numbers and non-symbolic representations, discriminating

the size of two numbers or two arrays of objects, and identifying the position of a number on a number line). Evidence from professional development programs and curricular adaptations suggests a clear link between the development of such skills and children's math achievement (e.g., Clements & Sarama, 2008; Clements, Sarama, Layzer, Unlu, & Fesler, 2020; Clements et al., 2011).

Other studies that have focussed on training specific math and numerical skills have also provided positive evidence. For instance, although the link between non-symbolic and symbolic magnitude processing is heavily discussed (and controversial), Hyde, Khanum, and Spelke (2014) found that children who were trained on non-symbolic comparison—using a numerical discrimination task— showed improvements in symbolic mathematics when compared to a control group (see also Khanum, Hanif, Spelke, Berteletti, & Hyde, 2016). The numerical discrimination task is thought to activate the approximate number system (ANS), which contributes to understanding numerical magnitudes. There is also substantive evidence that number line estimation skills, which reflect children's knowledge of the relations among numbers and contribute to their understanding of addition and subtraction, may be effectively trained (for review see, Siegler, 2016). Other studies that have tackled several aspects simultaneously have also found positive effects. For instance, playing linear number board games such as "The Number Race", which involves numerical magnitude comparisons as well as arithmetic problems, has been found to improve number knowledge (Schacter & Jo, 2016; Schacter et al., 2016; see also, Siegler & Ramani, 2008). Similarly, Wilson, Dehaene, Dubois, and Fayol (2009) found that playing "The Number Race" increased children's numerical magnitude skills (see also, Räsänen, Salminen, Wilson, Aunio, & Dehaene, 2009; Sella, Tressoldi, Lucangeli, & Zorzi, 2016).

Noteworthy, there is evidence that training early numeracy skills may have more impact on children's math achievement than other programs that tackle domain-general skills (for a review, see Fischer, Moeller, Cress, & Nuerk, 2013). For instance, Ramani et al. (2020) investigated the benefits of training domain-specific and domain-general skills (WM-Updating) in kindergarten-aged children and found that only children who were assigned to the domain-specific condition (playing games that involved counting and numerical comparison) experienced improvements in math skills (measured as a latent variable that included both basic numerical skills as well as arithmetic skills). Similarly, Nemmi et al. (2016) found that a numeric training program based on number-line estimation induced math gains in 6-year-olds whereas a WM training did not.

Nonetheless, the effect sizes of numeracy training programs remain small, and the evidence of far-transfer effects (i.e., improvements in other than the trained skills) is mixed. For instance, Räsänen, et al. (2009; see also Sella et al., 2016) found improvements on trained tasks, or untrained but structurally equivalent tasks, after playing the Number Race game, but failed to observe changes in children's number line estimation skills. Similarly, Toll and Van Luit (2013) found that children who received training on more complex skills that included learning procedural and conceptual counting knowledge also improved on counting and arithmetic against a control group but did not outperform their peers in number line estimation performance.

Furthermore, similar methodological issues to those mentioned above (with regards to WM training programs) also arise in the context of numeracy training studies (for an extensive discussion, see Simms, McKeaveney, Sloan, & Gilmore, 2019). Several problems that affect the strength of the evidence of training persist in the literature, such as the inclusion of no-contact control groups or the absence of control groups (c.f., Laski & Siegler,

2014; Nemmi et al., 2016; van der Ven et al., 2017; Ramani et al., 2017), the measurement of abilities through single tasks, using outcome tasks that are structurally aligned with trained tasks, the lack of delayed post-testing (c.f., Navarrete et al., 2018), and statistical power.

Current gaps in the literature

Although the number of WM and numeracy interventions for younger children has substantially increased during the last decade, there are still gaps in the literature that call for further investigation. For instance, most evidence of WM improvements after WM training relates to improvements on simple-span tasks or average performance on simple- and complex-span tasks that are weak indicators of WM (for a comprehensive review of this issue, see Shipstead et al., 2012). Therefore, more evidence is needed regarding improvements on WM tasks that reflect more adequately the complexity and functions of the child's WM capacity. In this study, we used a combination of complex span and updating tasks for both training and evaluation to ensure that we were targeting improvement in the executive component of WM.

One possible reason for the absence of far-transfer effects in WM training studies is that improved math performance requires more than improved WM capacity. Specifically, although it has been argued that additional guidance, instructions, or reinforcement are required to help participants apply their newly developed skills to novel tasks or situations (Dunning et al., 2013; Randall & Tyldesley, 2016; but see Shipstead et al., 2012, for a counter-argument), another possible reason is that WM training protocols are usually "number-free". Using non-numerical stimuli during WM training may limit transfer to mathematical or numerical skills. This is even more critical for children with MLD. It has been suggested that the misalignment of outcome measures with the training's content has a more deleterious effect on word problem solving outcomes for struggling learners than for typically developing learners (Fuchs, Fuchs, Seethaler, & Barnes, 2020). Indeed, a growing

number of studies call for WM training to be embedded within the typical activities in which benefits are needed. Whilst there is some evidence that this may be beneficial (e.g., Kroesbergen et al., 2014), we do not know whether the effects of such programs are sustained over time.

It is also unknown how WM and numeracy interventions affect the rate at which math skills develop. Specifically, whether intervention alters the time course of development or variability in the speed of improvement across children. Many cognitive capacities increase linearly at the early stages and then level-off over development, which may affect the interpretation of delayed or long-term effects. Furthermore, prolonged intervention periods may be more effective in improving skills that naturally develop over longer periods of time. Although WM can contribute to arithmetic fluency, there are additional factors that contribute to such fluency and rely on extended experience with numbers (e.g., arithmetic facts). These possibilities underscore the need for longer duration of training to investigate how the effects of intervention programs unfold over time and differ from the normative trend.

Furthermore, we do not know whether children with MLD or at-risk of MLD may benefit more from one or other type of treatment to improve their math skills. The majority of WM training studies have targeted typically developing children and children presenting behavioural disorders such as ADHD (for a meta-analysis, see Peijnenborgh et al., 2016). Among the studies that have specifically targeted children at-risk of MLD and those with low WM capacity, some of them have not included both WM and numeracy training protocols (e.g., Holmes & Gathercole, 2014; Roberts et al., 2016), others have considered socioeconomic status (SES) as the inclusion criterion (e.g., Ramani et al., 2017, 2020), and others have not included an adequate control group (e.g., Nemmi et al., 2016) or show methodological issues like those mentioned above.

Crucially, it is unknown whether the positive effects that have been reported in experimental or lab-controlled environments can be replicated in real learning contexts. Although some small-scale studies have implemented lab-based studies in the classroom (e.g., Holmes & Gathercole, 2014), the requirements and settings of these studies still mimic lab-controlled environments rather than the real learning context. For instance, Cogmed is usually administered by school personnel or clinical practitioners who have been trained by Cogmed coaches (Simons et al., 2016). The translation of this and similar programs requires individual support and monitoring, which is challenging in typical learning environments with limited curriculum time, varied teachers' dispositions and students' engagement.

The current study

We designed three different WM-Updating and numeracy game-based interventions (WM-only, Numeric-WM, and Numeric-only) that were computerised, auto-administered and implemented in a classroom environment as part of normally scheduled class activities for one year. Participants were first-grade children identified by schools as having difficulties in mathematics. The main research question was whether these training protocols improved WM-Updating capacity and a range of math and numeracy related abilities including non-symbolic numerical discrimination, number-line estimation skills, and abilities as measured by a standardized math achievement test. We expected all three interventions to produce facilitative effects in their respective domains, but given prior findings, we were uncertain whether these positive effects would be sustained once the intervention is discontinued and whether there would be any transfer to untrained skills (far-transfer effects). We were particularly interested in and expected that the Numeric-WM condition (in which numerical content was embedded in WM training) would produce the most improvement since alignment of outcome measures with the training's content seems particularly relevant for children with learning difficulties (Fuchs et al., 2020).

Method

Sample

Children were recruited from those attending the Learning Support Program for Mathematics (LSM)—a program that is offered to children who may not have the necessary fundamental skills in numeracy and basic arithmetic at formal school entry. Approximately, 5% - 6% of children entering Primary 1 in Singapore enrol in the LSM program, which aligns with the extant prevalence rates of children at risk of MLD (Devine et al., 2013; von Aster et al., 2007).

A total of 428 children in 63 primary schools across Singapore were recruited (M_{age} = 83.9 months, SD_{age} = 4.35 months; 41% female) from 569 children who were screened for WM capacity (Backward Digit Recall task; adapted from Pickering & Gathercole, 2001), math fluency (Math fluency subtest of the Wechsler Individual Achievement Test WIAT-III; Wechsler, 2009), and reading skills (letter reading task of the Word Reading subtest of the WRAT-4; Wilkinson & Robertson, 2006). A CONSORT diagram can be found in the Supplementary Material (p.2). Of the 428 children, 216 were enrolled in both LSM and the Language Support Programme (LSP)—this program is designed for students in need of support in English, the language of instruction. The vast majority of the recruited children (82%) had low working memory capacity. They scored at or below the 20th percentile of scores on the same WM task from a similar age group in a previous longitudinal study with Singaporean children (Lee et al., 2013).

The majority of the children were from families with a median monthly household income of $5,000, measured on an 8-point scale from below $2,000 to above $14,000 in $2,000 gradation. This compares to a population median of $9,023 (Singapore Department of Statistics, DOS, 2017). The sample included representation from the three major ethnicities in Singapore. Parent consent and child assent were obtained before data collection. Ethics

approval was obtained from the first author's university institutional review board. All children received small tokens of appreciation for their participation.

Children were matched on screening scores, and pseudo-randomly divided into eight groups based on a 2 (difficulties grouping: LSM versus LSM+LSP) x 4 (training condition: WM-only [WM], Numeric-WM [NWM], Numeric [NUM], versus active control group [AC]) full factorial design. To achieve a balanced distribution, a similar number of children from each classroom were assigned to each of the eight groups. An ANOVA revealed no differences in screening measures between treatment conditions (WM: F (3, 424) = .132; Math: F = .563; Reading: F = .105). A similar number of students in LSM and LSM+LSP were assigned to each treatment condition (ratios ranged between .96 and 1) and no differences between treatment conditions were observed in the number of children assigned to LSM and LSM+LSP (χ^2 (3) = .038 and χ^2 (3) = .00, respectively). The cumulative attrition rate for the study was 3.7% (1.2%, 1.6%, and .9% for the test conducted at the midpoint of the intervention, immediate post-test, and delayed post-test, respectively).

Materials

A battery of tasks to assess participants' early numeracy and mathematic skills, as well as their WM-Updating capacity was administered at each time-point (four time-points). The tasks were divided into three task sets and administered over 3 sessions. Each set took 40 to 45 minutes to administer. Task and set order were counterbalanced across participants.

WM-Updating capacity

Children's WM-Updating capacity was measured with four tasks (*Running Letters* task, *Keep Track* task, *N-Back* task, and *Complex Span* task). Across tasks, children provided their responses verbally while the examiner wrote their responses verbatim on a record form. A high score indicated better updating capacity. Furthermore, before each task and block of

stimuli, a trained experimenter explained the mechanics of the task or block with flashcards. All tasks were computerized and have been used in several studies tackling the association between executive functions and mathematics in young children (for a meta-analysis, see Friso-van den Bos et al., 2013).

The *Running Letters* task was based on the running span paradigm by Pollack, Johnson, and Knaft (1959, as cited in Morris & Jones, 1990). First, we assessed children's knowledge of the letters that were used in this task (B, F, K, H, M, Q, R, X). Then, children were shown a series of letters one at a time (1500 ms and 500 ms inter-stimulus interval) and were asked to recall the last letter (increased to two, three, and four in later blocks) at the end of each trial. Children were not told how many items were to be presented. This task comprised four blocks of trials; each block had three practice trials followed by six experimental trials. The task was discontinued if participants scored less than 3 correct trials per block. The dependent measure was the total number of letters recalled correctly, regardless of the sequence in which the letters were presented.

The *Keep Track* task was based on the keep track paradigm by Yntema (1963). First, children were shown five categories (letters, sports, jobs, colors, and animals) and the exemplars (4 pictures) in each to ensure that they knew to which category each picture belonged. Then, children were shown a series of pictures one at a time (1500 ms and 500 ms inter-stimulus interval) and were asked to recall the last stimulus shown in each category (increased to two, three, and four categories in later blocks) at the end of each trial. Children were not told how many pictures were to be presented. This task comprised four blocks of trials (each block had three practice trials followed by six experimental trials). The task was discontinued if participants scored less than 3 correct trials per block. The dependent measure was the total number of correct trials.

The *N-Back* task required children to decide whether each letter in a sequence matched the one that appeared *n* letters ago. First, we assessed children's knowledge of the letters that were used as stimuli (B, F, K, H, M, Q, R, X). Then, children were shown a list of letters one at a time (1500 ms and 500 ms inter-stimulus interval) and asked to indicate whether the letter that was presented matched the one that had appeared one letter ago (increased to two and three letters ago in later blocks). Verbal responses were required for positive matches only (a "Yes" response). This task comprised three blocks; each block had one practice list of 16 letters where five were positive matches (i.e., Yes response; for instance, B F K K F X… for block #1 [i.e., *1-back*], where the second K denotes the target) followed by three experimental lists of 16 letters (each list included five positive matches). A thirty-second pause was included between each list of letters. The distribution of positive matches in each list of letters varied across lists. The task was discontinued if participants failed to identify less than 10 (of a maximum 15) positive matches per block. The dependent measure was the total number of correct positive matches.

The *Complex Span* task was based on the short version of the Rotation task of Foster et al. (2015). Children first saw a letter presented either normally, or mirrored, which was rotated on its vertical axis. They were tasked to determine whether the rotated letter was presented normally or mirrored. Then, children were presented with an arrow pointing in one of eight directions. They were then asked to recall the direction of the arrows. The number of letter and arrow pairs ranged from 2 to 9 per trial, and children recalled the arrows in the order they were presented. This task comprised four blocks of trials; each block had three practice trials followed by six experimental trials. In block #1, children had to recall the last letter and arrow pair that was presented in each trial; in block #2, the last two pairs, and so on. The task was discontinued if participants scored less than 3 correct trials per block. The dependent measure was the total number of correct trials (arrows).

Early numeracy skills (number line estimation and numerical discrimination acuity)

We used two measures that have been used widely in studies that have investigated core underlying skills that support the acquisition of mathematical abilities in school-age children. For instance, poor number line estimation skills are thought to characterize children at risk of developing MLD in elementary school (Bull et al., 2021). Similarly, numerical discrimination acuity of children with mathematical difficulties and dyscalculia is significantly poorer than that of their typically achieving peers (Skagerlund & Träff, 2016).

In the *Number-Line* task, children were tasked to indicate on a line where a number at the top should go. Children were tasked to solve two different versions of the *Number-Line* task (0-10 and 0-100). On each problem, a number between 1 and 9 (and 1 and 99 in the 0-100 version) was presented along with a horizontal number-line in the middle of a computer screen with 0 at the left end and 10 at the right end (and 0 and 100 in the 0-100 version). Children completed four practice trials, after which the remaining numbers were presented, one at a time without feedback. All of the numbers from 1 to 9 were presented on the 0-to-10 task. Each number was presented twice to calculate an average positioning. In the 0-to-100 version, only the numbers 3, 4, 6, 8, 12, 14, 17, 18, 21, 24, 25, 29, 33, 39, 42, 48, 52, 57, 61, 64, 72, 76, 79, 81, 84, 88, 90, 94 and 96 were presented once. Number line estimation proficiency was based on the percentage of absolute error (PAE). The PAE was calculated as [(actual number estimated – target number presented)/scale of number-line] x 100. A lower PAE indicates better estimation skills.

In the *Numerical Discrimination* task, each trial consisted of the presentation of two arrays of dots on a computer screen which were described to the children as coins to be collected. The arrays were presented for 2 seconds, after which a screen appeared showing two question marks. The child was asked to indicate which display showed more coins by pressing a key corresponding to the side of the more numerous array. Feedback was provided

after each trial, and after every 20 trials a feedback screen was presented to encourage the child. The number of dots of each color in the array varied from 5 to 35, with each pair depicting a ratio difference of 0.91 (e.g., 10:11), 0.83 (e.g., 5:6), 0.77 (e.g., 7:9), 0.71 (e.g., 5:7), or 0.67 (e.g., 6:9). At each ratio level absolute difference between the stimuli pairings differed, e.g., at ratio 0.67, stimulus pairings were 6:9, 10:15, 12:18, 18:27, and 20:30. Five pairs at each ratio level were presented four times, the highest number appearing equally often in each color on each side. This resulted in a total of 100 trials, 20 at each ratio difference. All trials were area controlled, and to ensure children were responding on the basis of quantity and not dot size, individual item size was varied to ensure that items in the less numerous arrays were not always larger than those in the more numerous arrays. Accuracy (% of correct trials) was the dependent measure.

Mathematical achievement

Math skills were measured with three pen-and-paper tasks (*Numerical Operations, Math Problem Solving*, and *Math Fluency* subtests of the WIAT-III, Wechsler, 2009). The WIAT-III is an instrument frequently used to measure children's academic achievement and to identify math learning difficulties (see Schroeder, 2020, for a recent implementation). Each task measures a different aspect of the math abilities that children are expected to develop over the first year in school. The *Math Fluency* subtest consists of 48 sums each for Addition and Subtraction and requires children to complete as many sums as possible in each set within a minute. The raw score was the total number of addition and subtraction problems solved correctly within the time limit. The *Numerical Operations* subtest assesses skills in identifying and writing numbers, rote counting, number production, and solving written calculation problems and simple equations, drawing from the basic operations of addition, subtraction, multiplication, and division. The *Math Problem Solving* subtest is a verbal problem-solving test that (for this age group) measures the ability to count, identify geometric

shapes, and solve single and multi-step word problems with the aid of visual cues. For both tasks, testing was discontinued after the child made 6 consecutive incorrect responses, and raw scores were used in the analyses.

Intervention (training)

Four adaptive computerized game-based training protocols were developed (Working Memory training—WM, Numeric WM training—NWM, Numeric training—NUM, and Active Control—AC). A detailed description of the games is provided in the Supplementary Material (p.3). In the WM training, children were tasked with a series of activities that required storing, updating, and retrieving information according to the four WM-Updating paradigms previously described under WM-Updating capacity tasks. In the NWM training children were tasked with similar activities, but we used numerical stimuli. In both tasks, difficulty increased as training progressed. This was effected by increasing the number of items that children had to recall from one to four as well as shifting from easy-to-encode items (e.g., fruits in the WM training and number symbols in the NWM training) to difficult-to-encode items (e.g., aliens in the WM training and non-symbolic magnitudes in the NWM training). NUM training had no mnemonic component. Activities included transcoding between symbolic and non-symbolic numerical magnitudes, number line estimation, numerical magnitude comparison, and addition and subtraction of small numbers. Difficulty in this game was manipulated as a function of the size of numbers or the ratio between numbers in the magnitude comparison activities. Children assigned to the AC condition were exposed to the same stimuli that were shown in the WM game, but they were not tasked to store and recall information (the recall phase was not included).

All of the games were adaptive to the users' performance and did not require one-to-one administration or the teachers to supervise gameplay. Thus, the number of trials within each game (game progression or treatment exposure) varied per session. Furthermore, the

number of sessions could also vary as a function of teaching needs given the scale and length of implementation. For these reasons, the percentage of game completed or game progression reflects more accurately the dosage or exposure to treatment than then the number of sessions.

At the end of each trial (independently of training treatment), children received feedback in the form of badges and virtual gold coins that served to improve non-playable aspects of the game that related to each participant's progression in the game (e.g., spaceship improvements). During the first training session, children were given the game instructions that were embedded into a storyline. Each of the treatment conditions included four themes and an overarching meta-story to maintain engagement. All children cycled through the themes during the training. The training games were played on identical 7-inch touchscreen tablets with children's responses being recorded on the tablet. We collected data regarding the overall amount of experience they had on the game (i.e., game progression or percentage of game completed), and the time children played the game (i.e., gaming time).

Procedure

Children were tested individually four times: pre-test, mid-intervention, immediate post-test, and delayed post-test (testers were blind with respect to the treatment conditions). Pre-test and immediate post-test corresponded to the beginning and end of the computerized training, respectively. The gap between data points was approximately six months, and data collection for each time point spanned about 2 months. The intervention was rolled out in batches immediately after the pre-test. We planned the intervention to be administered twice or thrice a week (10-15 minutes per session) for 40 weeks during the LSM sessions, which would be equivalent to approximately 24 hours of training. Demographic information was collected on entering the study.

Analytical approach

We used a multi-group latent growth curve model (MG-LGC) to estimate the effects of treatment on the longitudinal trajectory and growth rate of children's WM-Updating capacity, early numeracy, and math skills. Data from the training and active control groups were analysed separately though simultaneously—i.e., the model estimated four different curves corresponding to the four groups. For each growth curve, the intercept corresponded to performance on the modelled ability at the beginning of the intervention. The slope related to the growth rate (per month) of the same ability from the beginning to six months after the termination of training (18-20 months). We adopted the approach described in Muthén and Curran (1997) to test the effectiveness of the various training treatments. First, we investigated the shape of the growth curve for each of the four conditions. Then, we modelled an MG-LGC for each outcome variable where the growth estimates (intercept, slope, and their corresponding variances and covariances) were constrained to equality across conditions. Muthén and Curran (1997) referred to these as normative estimates. Simultaneously, we estimated two additional linear growth factors for each of the three training groups. These additional growth factors were freely estimated across training groups (i.e., allowed to vary). The first additional growth factor (G1) measured growth from the beginning of the intervention to the immediate post-test; the second additional growth factor (G2) measured growth from the immediate post-test to the delayed post-test. The means of G1 and G2 correspond to growth rate differences with regards to the normative growth rate (immediate and delayed or long-term treatment effects, respectively). Although test of treatment effects can be accomplished by testing for changes in model fit using a conventional multiple group approach in which parameters are constrained then freed across groups, the current approach allows for more flexibility in modelling. Specifically, using an additional growth factor to estimate the treatment effect allows for characterization of the

shape of that effect, the amount of variation in that effect across individuals, and testing of covariates that influence that effect.

Because the study was implemented during regular class time, differences in training exposure with the games would likely affect the outcomes of the training. Thus, we included the percentage of game completed as a covariate that influenced the treatment effects G1 and G2. The covariance between percentage of game completed and the normative growth factors (intercept and linear and quadratic slopes) was fixed at zero since these growth factors reflect the normative trend for which no effect of game exposure is hypothesized.

WM-Updating measures and the math measures from the WIAT were modelled as two different latent variables[1]. Although latent variables offer advantages over manifest or observed variables in dealing with statistical assumptions and in adjusting for measurement error, they also require strong assumptions about the invariance of the factor structure over time when a growth model is specified. If the underlying factors are substantially different, then there is no basis for interpreting observed differences or modelling growth over time. Thus, we conducted a measurement invariance analysis for each latent construct. Factor scores derived from these analyses were used as outcome variables in the corresponding growth model. Preliminary analyses to investigate the invariance of the latent constructs (WM-Updating and Math) over time as well as those regarding the longitudinal trajectory of each treatment group for the various outcome measures are shown in Supplementary Material (p.13).

Data were screened for outliers (< 2% per variable). We replaced scores that were more than 3 SD from the sample mean with values computed at 3 SD from the mean. Missing data, corresponding to absent participants and other administration issues during testing

[1] We did not find an underlying factor representing variability in the early numeracy tasks. Indeed, an underlying factor reflecting performance on both number-line tasks (or a composite) was only appropriate for the last time point (correlations ranged between .20 at entry to the study and .40 at the delayed post-test).

(ranged between 0% and 1.3% across outcome measures) were treated as missing at random. All descriptive and inferential statistical analyses were estimated using Mplus (Version 8.6; Muthén & Muthén, 1998–2017). The TSCORES option in Mplus was used to account for time-varying observations since the intervention was rolled out in batches and each time point of data collection spread over several weeks. This method only provided relative fit indices (the Akaike information criterion—AIC, and the Bayesian information criterion—BIC). Differences in AIC and BIC values afford comparisons of models with differing numbers of parameters. Smaller BIC ($\Delta > 2$) and AIC ($\Delta > 9$) suggest a better model fit (Raftery, 1995). In the current study, we used BIC because this index is more consistent in selecting the true model when the true model is a candidate (Vrieze, 2012). We used a robust maximum likelihood estimator (MLR), with standard errors that are robust in relation to non-normality and non-independence of observations. We also applied corrected standard errors to control for the nesting of students within schools (TYPE = COMPLEX option in Mplus). Effect sizes (Hedges' *g*) regarding the immediate and long-term effects of each training protocol were estimated as the difference in slope coefficients between the Active Control group and each of the training groups (i.e., additional growth factor G1 or G2) divided by the pooled standard deviation at Time 1 (Feingold, 2009).

Results

Means and standard deviations for all the outcome variables and time points (per treatment condition) are shown in Table S-1 (Supplementary Material, p. 10). Given the results of the single-group analyses (see Supplementary Material, p.16), and that the mean of the quadratic term was significantly different from zero for all the outcome measures and treatment groups, a quadratic normative curve was specified. The model that was used to

investigate treatment effects during the intervention period and long-term effects six months later (G1 and G2, respectively) is shown in Figure 1.

FIGURE 1

Note that this model is a conditional model that reflects the intervention effects in real-life contexts—where the training is auto-administered and adaptive, as game exposure *(% of game completed*) may vary across children. Indeed, we also found significant differences in game progression across training conditions—game progression in the WM and NWM training games was substantially smaller (see Figure 2, left panel). Half of the children assigned to these training protocols barely completed 20% of the game. Given that the median gaming time was similar across training conditions (see Figure 2, right panel), such limited game progression could be due to differences in game difficulty, or how challenging and cognitively demanding the WM and NWM training were for children in the current sample.

FIGURE 2

There were substantial differences in game progression among the three treatment conditions. Because we were not interested only in comparing treatment effects across training conditions but whether there were differences between each treatment condition and the AC condition (i.e., whether children benefited from any of the treatments), the effectiveness of training was estimated conditional on the percentage of game completed within each treatment condition. Percentage of game completed was entered into the model as a covariate for the treatment effects. Because of the differences in game progression, the estimated treatment effects for the various conditions were, in essence, the treatment effects at

the average rate of progression in each condition (20% for WM and NWM, and 60% for NUM).

Treatment effects

With these considerations in mind, we first investigated whether differences between treatment conditions existed at the beginning of the intervention. To that end, we constrained the initial status to be equal across treatment conditions, which did not lead to a deterioration in model fit. Model fit indices of this restrictive model were smaller than those of the original (unrestrictive) model for all treatment conditions (see Table 1, top panel; note that the unrestricted and restricted model are nested models). This means that no differences between treatment conditions existed at the beginning of the intervention and that the pseudo-randomization worked appropriately. Parameter estimates of this restrictive model are shown in Table 1 (bottom panel). Although our sample included children with two different profiles of difficulties (i.e., children receiving math support vs those receiving both math and language support), they did not benefit differently from the trainings. For clarity, the relevant coefficients have not been included in Table 1.

TABLE 1

In line with findings from the single-group analyses, this model showed that there was significant growth over time, but also a significant deceleration for most of the outcome measures (the only exception being number-line estimation 0-100). The model revealed that the magnitude of the immediate treatment effect (G1) varied as a function of children's experience with the games (*% of game completed*) for most of the outcome variables (the only exception being number-line estimation 0-100). Children who had more experience with the

games showed larger growth rates. This was also evident in children assigned to the WM and NWM training conditions (even though they produced no significant overall treatment effect). In other words, the WM-Updating growth rates over the intervention period were larger for children who progressed further in the intervention than for those progressed less far, when compared to growth rates in the AC condition. Similarly, children who completed a larger percentage of the NWM training benefited more from the training than those who completed less of the training—for NWM, this pattern of finding applied to math, number-line estimation, and numerical discrimination skills.

A significant (immediate) treatment effect was found concerning both number-line estimation tasks. Children in the NUM condition showed steeper (negative) growth (i.e., the percentage of absolute error decreased) in both tasks. Gains corresponding to the 0-to-10 task sustained over time. The finding that G2 was not significant indicates that children showed similar growth to those in the AC group once the intervention was over. Because children in the NUM training condition showed greater improvement at the end of intervention, the finding shows that children in the AC group did not catch up with those in the NUM training. In contrast, the positive coefficient of G2 in the 0-to-100 task indicates that the effect of the intervention vanished. Children in the AC condition showed a steeper decrease (i.e., smaller PAE) than those assigned to the NUM training once the intervention was over. Curiously, children in the NWM condition also exhibited a deterioration in their accuracy on the 0-to-100 estimation task at the delayed post-test. Unlike children in the NUM condition, these children did not significantly benefit from training at the immediate post-test. Thus, the findings suggest that children in the NWM condition performed worse than those in the AC condition at delayed post-test. The analysis also revealed that children in the NWM training showed steeper growth in numerical discrimination skills once the intervention was discontinued—long-term effects.

Conditional treatment effects

Because the overall progression in the WM and NWM conditions were low, we conducted a further set of analyses to estimate the effectiveness of treatment (as well as the treatment potential) when children were able to progress further in the intervention. Rather than simply splitting the sample into smaller groups and focussing on children who completed a specific percentage of the game, we estimated the conditional effect of the covariate at, approximately, the upper interquartile of game experienced for each training condition (i.e., the covariate was centered at 30% for WM and NWM, and 80% for NUM). This approach has the benefit of utilising the full sample and it is structurally the same model as that reported earlier. Parameter estimates of the conditional analyses are shown in Table 2.

TABLE 2

At higher percentages of game completed (upper interquartile, 30% for WM and NWM, and 80% for NUM), the benefits of the NUM training are more evident. For instance, the immediate gains (G1) in both number-line estimation tasks remained once the intervention was over (i.e., the non-significant G2 effect indicates that the differences observed at the immediate post-test remained at the delayed post-test). In addition, children assigned to the NUM training condition also showed larger growth rates in math skills that were already evident at the immediate post-test and persisted once the intervention was discontinued—i.e., the non-significant long-term effect (G2) indicates that differences between NUM and AC groups did not vary from the immediate post-test to the delayed post-test. Long-term treatment effects regarding the numerical discrimination task were also observed in children assigned to the NUM training.

In contrast, no evidence that higher percentages of game completed in the WM training affected the treatment potential was found. Similarly, the treatment potential of the NWM training did not change substantially at higher percentages of game completed. Only a positive effect was found regarding number line estimation skills (0-10 task).

Discussion

Although the number of WM and numeracy training studies has increased exponentially during the past decade, it is not clear whether children with MLD or at-risk of MLD can benefit from them. Arguably, the idea that WM training can effectively improve WM capacity and (indirectly) contribute to mathematical performance is attractive. Indeed, some studies have suggested that the effect of WM training may be more evident for those with lower WM capacity and learning disabilities[2] (Klingberg, 2010; Klingberg, et al., 2002; Weicker et al., 2016). Nonetheless, our results showed no near- or far-transfer effects for WM training in children at-risk of MLD.

Since difficulties in math can stem from domain-general (e.g., WM, attention), domain-specific (e.g., numerical magnitude processing), or a combination of both types of factors, it has been suggested that embedding WM and numeracy training may be more beneficial for struggling learners (Fuchs et al., 2020). In the current study, the NWM training protocol embedded numerical activities within the context of WM training. For instance, it included a game that required participants to estimate the position of a number on a number line while storing (and recalling) the shapes wherein each number was encapsulated. We

[2] A post-hoc analysis revealed an aptitude-by-treatment interaction, hence, suggesting that children with better cognitive skills at screening benefitted more from the WM training. The magnitude of the treatment effect regarding WM-Updating skills (differences in the growth rate of WM-Updating skills between children assigned to the WM training and those assigned to the AC group) increased as a function of the child's math scores at screening. Children with better math skills at screening had larger treatment effects at the immediate post-test (b= .01, p < .01). These children also experienced steeper declines at the delayed post-test (b= -.02, p < .05), which suggests that potential WM training effects decay once the intervention is discontinued. Similarly, the magnitude of the differences in WM-Updating skills (between children in the AC group and those assigned to the NWM treatment) at the immediate post-test varied as a function of math and WM capacity at screening—a larger immediate effect for children with better math and WM skills (b= .01, p < .05 and b= .03, p < .05, respectively).

found that the NWM training program induced long-term gains in numerical discrimination acuity and that children who had more experience with the games also showed better number-line estimation skills during the training. Nonetheless, these effects were also identified in children allocated to the numeric training (NUM). Because the NWM training did not produce any WM-Updating improvements, it is more likely that the improvement results from the numeracy component of the training[3].

In this sense, our findings are consistent with other studies that have found that numeric training may benefit children at risk of MLD to a larger extent (e.g., Nemmi, et al., 2016). We found that children allocated to the NUM training showed better performance than those in the AC condition in their number-line estimation skills. The conditional effect also supported the treatment potential. Those who had more experience with the games also showed better math skills and numerical discrimination acuity. It should be noted that there are overlaps in the content of training and the outcome measures. Both number line estimation and numerical discrimination were part of the NUM training. This specificity of training effects has also been observed in other numerical training studies that have focused on particular aspects of children's math development (e.g., Räsänen, et al., 2009; Sella et al., 2016, Toll & Van Luit, 2013). Indeed, improvements on trained tasks are the norm rather than the exception and may be considered a sign of treatment potential. Noteworthy, as discussed below, children who completed 80% of the NUM training showed improvements in the math latent factor that reflected other than the trained skills.

Our findings suggest that the efficacy of cognitively demanding interventions may be reduced if existing cognitive resources are very limited[4]. For a WM training program to be

[3] A post-hoc analysis revealed an aptitude-by-treatment interaction in which children with better math skills at screening showed a larger immediate effect regarding performance on number-line estimation (b= -.13, $p < .001$). This is in line with other studies that have found that initial math skills level may moderate the effectiveness of math interventions (Clarke et al., 2020)

[4] Note that this is a tentative explanation and that a sample with a wider range of WM abilities would be needed.

effective, children must work at maximum WM capacity (Klingberg, 2010), which means under conditions of increased cognitive load. Most of the children in our study had WM capacity below the 20th percentile. It is possible that though the intervention games are adaptive, the task demands may have overloaded the limited cognitive resources of these children. Thus, they were unable to benefit from the training. The limited game progression of children assigned to both WM training conditions, though not conclusive, is consistent with this interpretation and may illustrate how these children struggled with trials that involved difficult-to-phonologically-recode items such as aliens. These trials were presented when children had completed about 30% of the game. Whereas children were able to perform relatively well on span-3 and -4 trials when familiar items such as fruits were presented, they failed on span-1 and -2 trials when recalling involved non-familiar items (each alien included unique morphological features to support encoding in memory—curvy, spiky). Note also that both the WM training and WM outcome measures targeted WM-updating skills, which may trigger heavier computational demands than other simple-span WM tasks targeting verbal or visual short-term memory (e.g., forward digit and word recall), such as those included in Cogmed training. Indeed, it is not yet clear that simple span WM training improves other than short-term memory—which does not shed light on whether the child's WM resources increase effectively (Shipstead et al., 2012). For instance, Shavelson et al. (2008) used Cogmed training and found no evidence of "near-transfer" effects when complex WM span tasks were used.

It might be argued that the auto-administered protocol could be responsible of the limited training/dosage, hence, constraining our interpretation regarding the potential effect of WM trainings. Nonetheless, similar findings have been reported in other large-scale studies with children of similar characteristics and "adequate" training exposure. For instance, children in Roberts et al. (2016) were below the 15th percentile in a task measuring

WM capacity and were administered the specific dosage (number of sessions) of Cogmed training. Nonetheless, no improvement was observed. This also aligns with findings from two meta-analyses on the effectiveness of WM training in children with low WM capacity. For instance, Peijnenborgh et al. (2016) focussed on ADHD and found that WM training programs seemed more effective for older children than those in the current study (<10 years old). Recently, Aksayli et al. (2019) extended substantially the corpus of that meta-analysis and found that WM-training programs such as Cogmed do not exert stronger benefits to low-WM individuals. Indeed, Nemmi et al. (2016) found in a computerized WM- and numeric-training with 6-year-olds that larger gains related to children with better math achievement and better WM capacity.

Although not reported in the Results section, we did not find that children with different support needs (math vs. math and language) benefitted differently from the trainings. Nonetheless, children who were receiving both math and language support had poorer WM-updating capacity and math and early numeracy skills. Whereas language difficulties could affect performance on tasks that involved a verbal component (e.g., Math Problem Solving), children in the LSM+LSP group (those receiving both language and math support) also showed poorer performances on tasks measuring other basic numeracy skills that do not rely on language proficiency. For instance, in the Numerical Discrimination Task, children were tasked to simply indicate which of two arrays of dots was larger (i.e., non-symbolic magnitude comparison). This finding suggests that children in the LSM+LSP group have more entrenched difficulties that go beyond their cognitive and language capacities, which aligns with some research on the comorbidity of reading and math skills that has suggested a multi-deficit pattern that involves shared etiological factors affecting both verbal and non-verbal aspects of the child's development (Cirino, Child, & Macdonald, 2018; Pennington, 2006). This finding also adds to extant research on MLD and suggests deficits

that go beyond math facts and simple computation, which are core aspects to determine MLD prevalence at entry to formal school in many countries.

Are training effects sustained over time?

We found long-term effects with regard to number-line tasks and math in children who had more experience with the games; differences that were observed at the immediate post-test were still evident six months later. This suggests that the intervention contributed to children's refinement of the numerical system to a larger extent than the regular classroom experience. A different pattern was found regarding numerical discrimination acuity; children who completed a higher percentage of the NUM games only started showing better numerical discrimination acuity than children in the AC condition only during the six months after the intervention was finished (i.e., G2 was significant). Although we cannot conclude that such differences are exclusively due to children in the NUM training condition having more experience with the numerical discrimination task during the intervention period (as contrasted to a generalised improvement in numerical discrimination abilities per se), this finding suggests that such refinement of the approximate number system (ANS) only emerges after substantial practice or experience with numbers. Note that children in the NWM training also showed steeper growth rates during the post-intervention period.

It is worth mentioning that the current study uncovered patterns that influence the interpretation of sustainability in the context of studies that have investigated long-term effects without a control group as well as among those studies that have considered shorter intervention periods. For instance, we found that the normative trend of all of the outcome measures included a growth deceleration (non-linear growth), which suggests that failing to identify long-term effects in studies that have not included a control group may simply reflect non-linearity of growth due to task artifacts or limitations in cognitive abilities that would also affect the normative trajectory.

Similarly, our study calls attention to findings from studies that have considered shorter intervention periods or have not investigated long-term effects. For instance, at the median value of *percentage of game completed*, the immediate effect of the NUM training on the 0-to-100 Number Line was significant, but the long-term effect was not—children in the AC condition still showed some growth during that period and performed at a similar level than those in the NUM training six months after the intervention. Similarly, the intervention effects regarding children's numerical discrimination skills only emerged in the long-term.

Classroom implementations of WM and numeracy training programs are challenging

Although the entrenched cognitive deficits of children with or at-risk of MLD may pose constraints to training programs that are cognitively demanding per se and require effortful control and attention, one of the reasons for a modest effect size is the self-contained nature of the intervention. One of the design constraints for our intervention is to move away from designs that require a high degree of supervision and monitoring, which is difficult to achieve in most typical school settings. Teachers in the current study were not asked to closely monitor each student's progression and it is feasible that some children did not engage with the games—children with weak WM capacity are usually inattentive and distractible (Alloway, Gathercole, Kirkwood, & Elliott, 2009). Ensuring that children are working at capacity while ensuring that the degree of monitoring is manageable within a classroom setting is a balance that still requires further experimentation. In this vein, it should be noted that the interventions used in the current study should be viewed more as supplemental to regular math instruction rather than replacing math instruction—in contrast to more intensive approaches that consider curriculum adaptations/accommodations.

It is worth mentioning that it is unlikely that teachers contributed (directly or indirectly) to the low percentages of game completed in children assigned to the WM and NWM protocols. As much as possible, we aimed at getting a balanced distribution of training

conditions within each class. That is, the same teacher was handling several tablets that corresponded to the four different conditions. Teachers were informed about the nature of the study, but they did not know who was assigned to each condition. They were informed on how to switch on and off the tablet and log on/off the game, but the implementation of the training did not require teachers to guide the students through the games or to monitor the students. Furthermore, the gaming time (see Figure 2, Results section) was similar across treatment conditions—independently of difficulty.

It should be noted that the study also suffered from poor compliance, which is a major threat to obtaining statistical power to detect intervention effects (Jo, 2002). On average, children were exposed to the games for about 6 hours (range= .6 – 24.2). Among the reasons for poor compliance, teachers mentioned children's frequent absenteeism, limited curriculum time, the need to prioritize curriculum time against gaming time, and that some children were not keen to play. The limited gaming time could affect children's experience with the games (proportion of game completed) as well as the magnitude of the effects reported here. Note also that finding smaller effect sizes is not uncommon when studies are scaled up to a population level.

Although these limitations may have attenuated the effect sizes, they enhanced the ecological validity of the study. As far as we know, ours is the first large-scale study that has particularly aimed at adapting and testing an auto-administered training program for children at-risk of MLD, and the fact that the study suffered from poor compliance indicates that implementations in real classroom settings are complex and may be affected by variables not directly related to the training itself. The fact that significant treatment effects were found despite these limitations suggests that such training should be further pursued. Note that the children in the current study presented a complex aetiology that included difficulties in math and language skills as well as very limited cognitive capacity, which can pose constraints to

any kind of training. Our findings also call attention to future interventions and training studies regarding their translation into tools to support teaching efforts and how to overcome the inherent limitations of real classroom settings.

Future studies should also investigate whether other types of math skills are more responsive to intervention. Although we used a measurement model that reflected the child's math ability and included arithmetic fluency as well as math problem solving skills, it is feasible that procedural skills such as arithmetic are more responsive than other skills that require creating both a mathematical model and a situation model (e.g., word problem solving). Similarly, whether effects would be similar for other aspects and types of WM (and for older children) is an open question that warrants further investigation. Providing a more balanced focus on verbal and visuo-spatial memory may be a good way to move forward.

References

Aksayli, N. D., Sala, G., & Gobet, F. (2019). The cognitive and academic benefits of Cogmed: A meta-analysis. *Educational Research Revie*w, 27, 229–243. http://dx.doi.org/10.1016/j.edurev.2019.04.003

Alloway, T. P., Gathercole, S. E., Kirkwood, H., & Elliott, J. (2009). The cognitive and behavioral characteristics of children with low working memory. *Child Development, 80*(2), 606-621. http://dx.doi.org/10.1111/j.1467-8624.2009.01282.x

Ang, S. Y., Lee, K., Cheam, F., Poon, K., & Koh, J. (2015). Updating and working memory training: Immediate improvement, long-term maintenance, and generalisability to non-trained tasks. *Journal of Applied Research in Memory and Cognition, 4*(2), 121-128. http://dx.doi.org/10.1016/j.jarmac.2015.03.001

Au, J., Gibson, B. C., Bunarjo, K., Buschkuehl, M., & Jaeggi, S. M. (2020). Quantifying the difference between active and passive control groups in cognitive interventions using two meta-analytical approaches. *Journal of Cognitive Enhancement, 4*(2), 192–210. http://dx.doi.org/10.1007/s41465-020-00164-6

Baddeley, A. (2012). Working memory: theories, models, and controversies. *Annual Review of Psychology, 63,* 1-29. http://dx.doi.org/10.1146/annurev-psych-120710-100422

Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences, 4*(11): 417-423. http://dx.doi.org/10.1016/S1364-6613(00)01538-2

Best, J. R., & Miller, P. H. (2010). A developmental perspective on executive function. *Child Development, 81*(6), 1641-1660. http://dx.doi.org/10.1111/j.1467-8624.2010.01499.x

Bull, R., & Lee, K. (2014). Executive functioning and mathematics achievement. *Child Development Perspectives, 8*(1), 36-41. http://dx.doi.org/10.1111/cdep.12059

Bull, R., Lee, K., & Múñez, D. (2021). Numerical magnitude understanding in kindergartners: a specific and sensitive predictor of later mathematical difficulties? *Journal of Educational Psychology*, *113*(5), 911-928. https://doi.org/10.1037/edu0000640

Butterworth, B., & Laurillard, D. (2010). Low numeracy and dyscalculia: identification and intervention. *ZDM, 42*(6), 527-539. http://dx.doi.org/10.1007/s11858-010-0267-4

Cirino, P. T., Child, A. E., & Macdonald, K. T. (2018). Longitudinal predictors of the overlap between reading and math skills. *Contemporary Educational Psychology, 54*, 99-111. http://dx.doi.org/10.1016/j.cedpsych.2018.06.002

Clarke, B. et al., (2020) Examining the efficacy of a kindergarten mathematics intervention by group size and initial skill: Implications for practice and policy. *The Elementary School Journal, 121*(1), 125-153. http://dx.doi.org/10.1086/710041

Clements, D. H., & Sarama, J. (2008). Experimental evaluation of the effects of a research-based preschool mathematics curriculum. *American Educational Research Journal, 45*(2), 443–494. https://doi.org /10.3102/0002831207312908

Clements, D. H., Sarama, J., Layzer, C., Unlu, F., & Fesler, L. (2020). Effects on mathematics and executive function of a mathematics and play intervention versus mathematics alone. *Journal for Research in Mathematics Education, 51*(3), 301–333. http://dx.doi.org/10.5951/jresemtheduc-2019-0069

Clements, D. H., Sarama, J., Spitler, M. E., Lange, A. A., & Wolfe, C. B. (2011). Mathematics learned by young children in an intervention based on learning trajectories: A large-scale cluster randomized trial. *Journal for Research in Mathematics Education, 42*(2), 127–166. http://dx.doi.org/10.5951/jresematheduc.42.2.0127

Cogmed, J. M. (2015). Cogmed Working Memory Training. *Behavioural Interventions to Remediate Learning Disorders: A Technical Report, 20*.

Cowan, N. (2016). Working memory maturation: Can we get at the essence of cognitive growth? *Perspectives on Psychological Science, 11*(2), 239-264. http://dx.doi.org/10.1177/1745691615621279

Devine, A., Soltesz, F., Nobes, A., Goswami, U., & Szűcs, D. (2013). Gender differences in developmental dyscalculia depend on diagnostic criteria. *Learning and Instruction, 27*, 31-39. https://doi.org/10.1016/j.learnintruc.2013.02.004

Dunning, D. L., Holmes, J., & Gathercole, S. E. (2013). Does working memory training lead to generalized improvements in children with low working memory? A randomized controlled trial. *Developmental Science, 16*(6), 915-925. http://dx.doi.org/10.1111/desc.12068

Feingold, A. (2009). Effect sizes for growth-modeling analysis for controlled clinical trials in the same metric as for classical analysis. *Psychological Methods, 14*(1), 43-53. http://dx.doi.org/10.1037/a0014699

Fischer, U., & Moeller, K., Cress, U., & Nuerk, H. C. (2013). Interventions supporting children's mthematics school success: A meta-analytic review. *European Psychologist. 18*. 89-113.

Foster, J. L., Shipstead, Z., Harrison, T. L., Hicks, K. L., Redick, T. S., & Engle, R. W. (2015). Shortened complex span tasks can reliably measure working memory capacity. *Memory & Cognition, 43*(2), 226-236. http://dx.doi.org/10.3758/s13421-014-0461-7

Friso-Van den Bos, I., Van der Ven, S. H., Kroesbergen, E. H., & Van Luit, J. E. (2013). Working memory and mathematics in primary school children: A meta-analysis.

Educational Research Review, 10, 29-44. http://dx.doi.org/10.1016/j.edurev.2013.05.003

Fuchs, L. S., Compton, D. L., Fuchs, D., Paulsen, K., Bryant, J. D., & Hamlett, C. L. (2005). The prevention, identification, and cognitive determinants of math difficulty. *Journal of Educational Psychology, 97*(3), 493-513. http://dx.doi.org/10.1037/0022-0663.97.3.493

Fuchs, L. S., Schumacher, R. F., Long, J., Namkung, J., Hamlett, C. L., Cirino, P. T., ... & Changas, P. (2013). Improving at-risk learners' understanding of fractions. *Journal of Educational Psychology, 105*(3), 683-700. http://dx.doi.org/10.1037/a0032446

Fuchs, L., Fuchs, D., Seethaler, P. M., & Barnes, M. A. (2020). Addressing the role of working memory in mathematical word-problem solving when designing intervention for struggling learners. *ZDM, 52*(1), 87–96. http://dx.doi.org/10.1007/s11858-019-01070-8

Geary, D. C., Hoard, M. K., Byrd-Craven, J., & DeSoto, M. C. (2004). Strategy choices in simple and complex addition: Contributions of working memory and counting knowledge for children with mathematical disability. *Journal of Experimental Child Psychology, 88*(2), 121-151. http://dx.doi.org/10.1016/j.jecp.2004.03.002

Gersten, R., Rolfhus, E., Clarke, B., Decker, L. E., Wilkins, C., & Dimino, J. (2015). Intervention for first graders with limited number knowledge: Large-scale replication of a randomized controlled trial. *American Educational Research Journal, 52*(3), 516-546. http://dx.doi.org/10.3102/0002831214565787

Gilmore, C. K., McCarthy, S. E., & Spelke, E. S. (2010). Non-symbolic arithmetic abilities and mathematics achievement in the first year of formal schooling. *Cognition, 115,* 394–406. http://dx.doi.org/10.1016/j.cognition.2010.02.002

Holmes, J., & Gathercole, S. E. (2014). Taking working memory training from the laboratory into schools. *Educational Psychology, 34*(4), 440–450. http://dx.doi.org/10.1080/01443410.2013.797338

Holmes, J., Gathercole, S. E., & Dunning, D. L. (2009). Adaptive training leads to sustained enhancement of poor working memory in children. *Developmental Science, 12*(4), F9-F15. http://dx.doi.org/10.1111/j.1467-7687.2009.00848.x

Hyde, D. C., Khanum, S., & Spelke, E. S. (2014). Brief non-symbolic, approximate number practice enhances subsequent exact symbolic arithmetic in children. *Cognition, 131*(1), 92-107. http://dx.doi.org/10.1016/j.cognition.2013.12.007

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Shah, P. (2011). Short- and long-term benefits of cognitive training. *Proceedings of the National Academy of Sciences of the United States of America, 108*(25), 10081–10086. http://dx.doi.org/10.1073/pnas.1103228108

Jo, B. (2002). Statistical power in randomized intervention studies with noncompliance. *Psychological Methods, 7*(2), 178. http://dx.doi.org/10.1037/1082-989X.7.2.178

Jones, J. S., Milton, F., Mostazir, M., & Adlam, A. R. (2019). The academic outcomes of working memory and metacognitive strategy training in children: A double‐blind randomized controlled trial. *Developmental Science,* e12870. http://dx.doi.org/10.1111/desc.12870

Karagiannakis, G., Baccaglini-Frank, A., & Papadatos, Y. (2014). Mathematical learning difficulties subtypes classification. *Frontiers in Human Neuroscience, 8*, 57. http://dx.doi.org/10.3389/fnhum.2014.00057

Karbach, J., Strobach, T., & Schubert, T. (2014). Adaptive working-memory training benefits reading, but not mathematics in middle childhood. *Child Neuropsychology, 21*:3, 285-301, DOI: 10.1080/09297049.2014.899336

Kassai, R., Futo, J., Demetrovics, Z., & Takacs, Z. (2019). A meta-analysis of the experimental evidence on the near- and far-transfer effects among children's executive function skills. *Psychological Bulletin. 145,* 165-188. http://dx.doi.org/10.1037/bul0000180

Khanum, S., Hanif, R., Spelke, E. S., Berteletti, I., & Hyde, D. C. (2016). Effects of non-symbolic approximate number practice on symbolic numerical abilities in Pakistani children. *PLoS ONE,11*(10), e0164436. http://dx.doi.org/10.1371/journal.pone.0164436

Klingberg, T. (2010). Training and plasticity of working memory. *Trends in Cognitive Sciences, 14*(7), 317-324. http://dx.doi.org/10.1016/j.tics.2010.05.002

Klingberg, T., Fernell, E., Olesen, P. J., Johnson, M., Gustafsson, P., …& Westerberg, H. (2005). Computerized training of working memory in children with ADHD-a randomized, controlled trial. *J. Am. Acad. Child Adolesc. Psychiatry 44,* 177–186. http://dx.doi.org/10.1097/00004583-200502000-00010

Klingberg, T., Forssberg, H., & Westerberg, H. (2002). Increased brain activity in frontal and parietal cortex underlies the development of visuospatial working memory capacity during childhood. *Journal of Cognitive Neuroscience, 14*(1), 1-10. http://dx.doi.org/10.1162/089892902317205276

Kroesbergen, E. H., van't Noordende, J. E., & Kolkman, M. E. (2014). Training working memory in kindergarten children: Effects on working memory and early numeracy. *Child Neuropsychology, 20*(1), 23-37. http://dx.doi.org/10.1080/09297049.2012.736483

Laski, E. V., & Siegler, R. S. (2014). Learning from number board games: You learn what you encode. *Developmental Psychology, 50*(3), 853. http://dx.doi.org/10.1037/a0034321

Lee, K., Bull, R., & Ho, R. M. (2013). Developmental differences in the structure of executive functions. *Child Development, 84*, 1933–1953.

Melby-Lervåg, M., & Hulme, C. (2013). Is working memory training effective? A meta-analytic review. *Developmental Psychology, 49*, 270–291. http://dx.doi.org/10.1037/a0028228

Morris, N., & Jones, D. M. (1990). Memory updating in working memory: The role of the central executive. *British Journal of Psychology, 81*(2), 111-121. http://dx.doi.org/10.1111/j.2044-8295.1990.tb02349.x

Muthén, B. O., & Curran, P. J. (1997). General longitudinal modeling of individual differences in experimental designs: a latent variable framework for analysis and power estimation. *Psychological Methods, 2*(4), 371. http://dx.doi.org/10.1037/1082-989X.2.4.371

Muthén, L.K. and Muthén, B.O. (1998-2017). *Mplus User's Guide*. Eighth Edition. Los Angeles, CA: Muthén & Muthén

Navarrete, J. A., Gómez, D. M., & Dartnell, P. (2018). Promoting preschoolers' numerical knowledge through spatial analogies: Numbers' spatial alignment influences its learning. *Contemporary Educational Psychology, 54*, 112-124. http://dx.doi.org/10.1016/j.cedpsych.2018.06.006

Nelson, G., & Powell, S. R. (2018). A systematic review of longitudinal studies of mathematics difficulty. *Journal of Learning Disabilities, 51(*6), 523-539.

Nemmi, F., Helander, E., Helenius, O., Almeida, R., Hassler, M., Räsänen, P., & Klingberg, T. (2016). Behavior and neuroimaging at baseline predict individual response to combined mathematical and working memory training in children. *Developmental Cognitive Neuroscience, 20*, 43–51. http://dx.doi.org/10.1016/j.dcn.2016.06.004

Nosworthy, N., Bugden, S., Archibald, L., Evans, B., & Ansari, D. (2013). A two-minute paper-and-pencil test of symbolic and nonsymbolic numerical magnitude processing explains variability in primary school children's arithmetic competence. *PLOS One, 8*(7), e67918. http://dx.doi.org/10.1371/journal.pone.0067918

Peijnenborgh, J. C., Hurks, P. M., Aldenkamp, A. P., Vles, J. S., & Hendriksen, J. G. (2016). Efficacy of working memory training in children and adolescents with learning disabilities: A review study and meta-analysis. *Neuropsychological Rehabilitation, 26*(5–6), 645–672. http://dx.doi.org/10.1080/09602011.2015.1026356

Peng, P., & Fuchs, D. (2016). A meta-analysis of working memory deficits in children with learning difficulties: Is there a difference between verbal domain and numerical domain? *Journal of Learning Disabilities, 49*(1), 3–20. http://dx.doi.org/10.1177/0022219414521667

Peng, P., Namkung, J., Barnes, M., & Sun, C. (2016). A meta-analysis of mathematics and working memory: Moderating effects of working memory domain, type of mathematics skill, and sample characteristics. *Journal of Educational Psychology, 108*(4), 455. http://dx.doi.org/10.1037/edu0000079

Pennington, B. F. (2006). From single to multiple deficit models of developmental disorders. *Cognition, 101*(2), 385-413. http://dx.doi.org/10.1016/j.cognition.2006.04.008

Pickering, S., & Gathercole, S. E. (2001). *Working memory test battery for children (WMTB-C)*. Psychological Corporation.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, 111-163. http://dx.doi.org/10.2307/271063

Ramani, G. B., Daubert, E. N., Lin, G. C., Kamarsu, S., Wodzinski, A., & Jaeggi, S. M. (2020). Racing dragons and remembering aliens: Benefits of playing number and

working memory games on kindergartners' numerical knowledge. *Developmental Science, 23*(4), e12908. http://dx.doi.org/10.1111/desc.12908

Ramani, G. B., Jaeggi, S. M., Daubert, E. N., & Buschkuehl, M. (2017). Domain-specific and domain-general training to improve kindergarten children's mathematics. *Journal of Numerical Cognition, 3*(2), 468-495. http://dx.doi.org/10.5964/jnc.v3i2.31

Randall, L., & Tyldesley, K. (2016). Evaluating the impact of working memory training programmes on children—A systematic review. *Educational and Child Psychology, 33*(1), 34–50.

Räsänen, P., Salminen, J., Wilson, A. J., Aunio, P., & Dehaene, S. (2009). Computer-assisted intervention for children with low numeracy skills. *Cognitive Development, 24*(4), 450-472. http://dx.doi.org/10.1016/j.cogdev.2009.09.003

Roberts, G., Quach, J., Spencer-Smith, M., Anderson, P. J., Gathercole, S., … & Wake, M. (2016). Academic outcomes 2 years after working memory training for children with low working memory: A randomized clinical trial. *JAMA Pediatrics, 170*(5), e154568–e154568. http://dx.doi.org/10.1001/jamapediatrics.2015.4568

Sala, G., & Gobet, F. (2017). Working memory training in typically developing children: A meta-analysis of the available evidence. *Developmental Psychology, 53*(4), 671–685. http://dx.doi.org/10.1037/dev0000265

Schacter, J., & Jo, B. (2016). Improving low-income pre-schoolers mathematics achievement with Math Shelf, a preschool tablet computer curriculum. *Computers in Human Behavior, 55*, 223–229.

Schacter, J., Shih, J., Allen, C. M., DeVaul, L., Adkins, A. B., Ito, T., & Jo, B. (2016). Math shelf: A randomized trial of a prekindergarten tablet number sense curriculum. *Early Education and Development, 27*(1), 74-88. http://dx.doi.org/10.1080/10409289.2015.1057462

Schmiedek, F., Hildebrandt, A., Lövdén, M., Wilhelm, O., & Lindenberger, U. (2009). Complex span versus updating tasks of working memory: the gap is not that deep. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(4), 1089. http://dx.doi.org/10.1037/a0015730

Schroeder, M., Drefs, M. A., & Zwiers, M. (2020). Comparing math LD diagnostic rates obtained using LDAC and DSM-5 criteria: Implications for the field. *Canadian Journal of School Psychology, 35*(3), 175-196. http://dx.doi.org/10.1177/0829573520915366

Schwaighofer, M., Fischer, F., & Bühner, M. (2015). Does working memory training transfer? A meta-analysis including training conditions as moderators. *Educational Psychologist, 50*(2), 138–166. http://dx.doi.org/10.1080/00461520.2015.1036274

Scionti, N., Cavallero, M., Zogmaister, C., & Marzocchi, G. M. (2020). Is cognitive training effective for improving executive functions in preschoolers? A systematic review and meta-analysis. *Frontiers in Psychology, 10*, 2812. http://dx.doi.org/10.3389/fpsyg.2019.02812

Sella, F., Tressoldi, P., Lucangeli, D., & Zorzi, M. (2016) Training numerical skills with the adaptive videogame "The Number Race": A randomized controlled trial on preschoolers. *Trends in Neuroscience and Education, 5*(1):20–9. http://dx.doi.org/10.1016/j.tine.2016.02.002

Shavelson, R. J., & Yuan, K. (2014). *On the Impact of computer cognitive training on working memory and fluid intelligence*. In Fostering Change in Institutions, Environments, and People (pp. 43-56). Routledge.

Shipstead, Z., Hicks, K. L., & Engle, R. W. (2012). Cogmed working memory training: Does the evidence support the claims? *Journal of Applied Research in Memory and Cognition, 1*, 185– 193. http://dx.doi.org/10.1016/j.jarmac.2012.06.003

Siegler, R. S. (2016). Magnitude knowledge: The common core of numerical development. *Developmental Science, 19*, 341-361. http://dx.doi.org/10.1111/desc.12395

Siegler, R. S., & Ramani, G. B. (2008). Playing linear numerical board games promotes low income children's numerical development. *Developmental Science, 11*, 655-661. http://dx.doi.org/10.1111/j.1467-7687.2008.00714.x

Simms, V., McKeaveney, C., Sloan, S., & Gilmore, C. (2019). *Interventions to improve mathematical achievement in primary school-aged children.* England, UK: Nuffield Foundation.

Simons, D. J., Boot, W. R., Charness, N., Gathercole, S. E., Chabris, C. F., Hambrick, D. Z., & Stine-Morrow, E. A. (2016). Do "brain-training" programs work? *Psychological Science in the Public Interest, 17*(3), 103-186. http://dx.doi.org/10.1177/1529100616661983

Skagerlund, K., and Träff, U. (2016). Number processing and heterogeneity of developmental dyscalculia: subtypes with different cognitive profiles and deficits. *Journal of Learning Disabilities, 49*, 36–50. http://dx.doi.org/10.1177/0022219414522707

Söderqvist, S., & Bergman‐Nutley, S. (2015). Working memory training is associated with long term attainments in math and reading. *Frontiers in Psychology, 6*, 1– 9. http://dx.doi.org/10.3389/fpsyg.2015.01711

St Clair-Thompson, H. L., & Gathercole, S. E. (2006). Executive functions and achievements in school: Shifting, updating, inhibition, and working memory. *Quarterly Journal of Experimental Psychology, 59*(4), 745-759. http://dx.doi.org/10.1080/17470210500162854

St Clair‐Thompson, H. L., Stevens, R., Hunt, A., & Bolder, E. (2010). Improving children's working memory and classroom performance. *Educational Psychology, 30*, 203– 219. http://dx.doi.org/10.1080/01443410903509259

Swanson, H. L., & Beebe-Frankenberger, M. (2004). The relationship between working memory and mathematical problem solving in children at risk and not at risk for serious math difficulties. *Journal of Educational Psychology, 96*(3), 471. http://dx.doi.org/10.1037/0022-0663.96.3.471

Swanson, H. L., Jerman, O., & Zheng, X. (2008). Growth in working memory and mathematical problem solving in children at risk and not at risk for serious math difficulties. *Journal of Educational Psychology, 100*(2), 343-379 . http://dx.doi.org/10.1037/0022-0663.100.2.343

Swanson, H. L., Olide, A. F., & Kong, J. E. (2018). Latent class analysis of children with math difficulties and/or math learning disabilities: Are there cognitive differences? *Journal of Educational Psychology, 110*(7), 931-951. http://dx.doi.org/10.1037/edu0000252

Szűcs, D. (2016). Subtypes and comorbidity in mathematical learning disabilities: Multidimensional study of verbal and visual memory processes is key to understanding. *Progress in Brain Research, 227*, 277-304.

Szűcs, D., Devine, A., Soltesz, F., Nobes, A., & Gabriel, F. (2013). Developmental dyscalculia is related to visuo-spatial memory and inhibition impairment. *Cortex, 49*(10), 2674-2688. http://dx.doi.org/10.1016/j.cortex.2013.06.007

Takacs, Z. K., & Kassai, R. (2019). The efficacy of different interventions to foster children's executive function skills: A series of meta-analyses. *Psychological Bulletin, 145*(7), 653-697. http://dx.doi.org/10.1037/bul0000195

Toll, S. W., & Van Luit, J. E. (2013). The development of early numeracy ability in kindergartners with limited working memory skills. *Learning and Individual Differences, 25*, 45-54. http://dx.doi.org/10.1016/j.lindif.2013.03.006

van der Ven, F., Segers, E., Takashima, A., & Verhoeven, L. (2017). Effects of a tablet game intervention on simple addition and subtraction fluency in first graders. *Computers in Human Behavior, 72*, 200-207. http://dx.doi.org/10.1016/j.chb.2017.02.031

Von Aster, M. G., & Shalev, R. S. (2007). Number development and developmental dyscalculia. *Developmental Medicine & Child Neurology, 49*(11), 868-873. http://dx.doi.org/10.1111/j.1469-8749.2007.00868.x

Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods, 17*(2), 228. http://dx.doi.org/10.1037/a0027127

Wang, C., Jaeggi, S. M., Yang, L., Zhang, T., He, X., Buschkuehl, M., & Zhang, Q. (2019). Narrowing the achievement gap in low-achieving children by targeted executive function training. *Journal of Applied Developmental Psychology, 63*, 87–95. http://dx.doi.org/10.1016/j.appdev.2019.06.002

Wass, S. V., Scerif, G., & Johnson, M. H. (2012). Training attentional control and working memory–Is younger, better? *Developmental Review 32,* 360–387. http://dx.doi.org/10.1016/j.dr.2012.07.001

Wechsler, D. (2009). *Wechsler Individual Achievement Test* (3rd ed). San Antonio, TX: Psychological Corporation.

Wilhelm, O., Hildebrandt, A. H., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *Frontiers in Psychology, 4*, 433. http://dx.doi.org/10.3389/fpsyg.2013.00433

Wilkinson, G. S., & Robertson, G. J. (2006). *Wide range achievement test*. Psychological Assessment Resources.

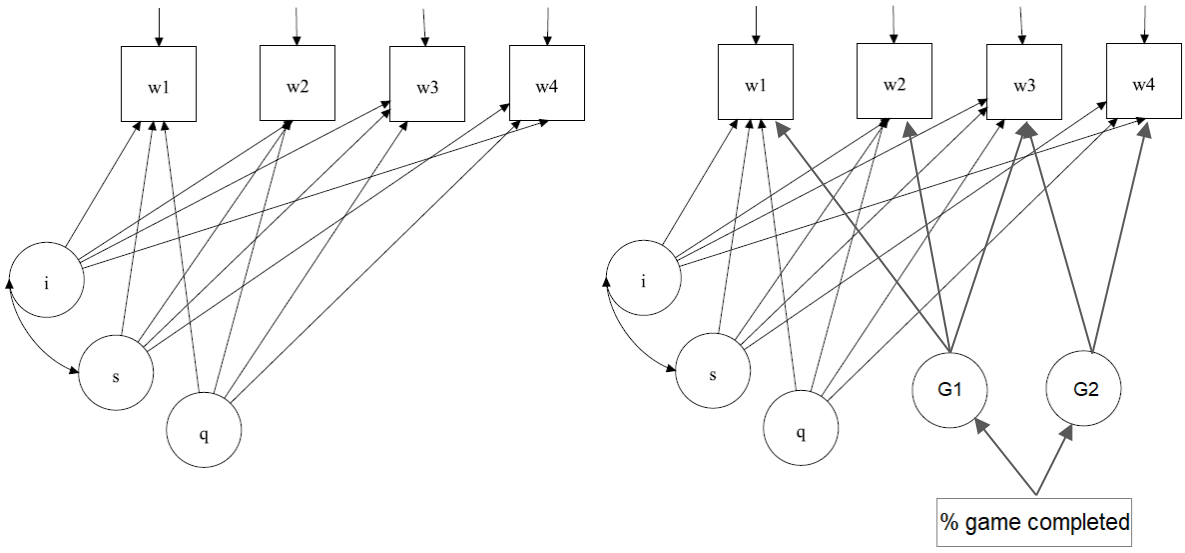
Wilson, A. J., Dehaene, S., Dubois, O., & Fayol, M. (2009). Effects of an adaptive game intervention on accessing number sense in low-socioeconomic-status kindergarten children. *Mind, Brain, and Education, 3*(4), 224–234. http://dx.doi.org/10.1111/j.1751-228X.2009.01075.x

Witt, M. (2011). School based working memory training: Preliminary finding of improvement in children's mathematical performance. *Advances in Cognitive Psychology, 7*, 7– 15. http://dx.doi.org/10.2478/v10053-008-0083-3

Yntema, D. B. (1963). Keeping track of several things at once. *Human factors, 5*(1), 7-17. http://dx.doi.org/10.1177/001872086300500102

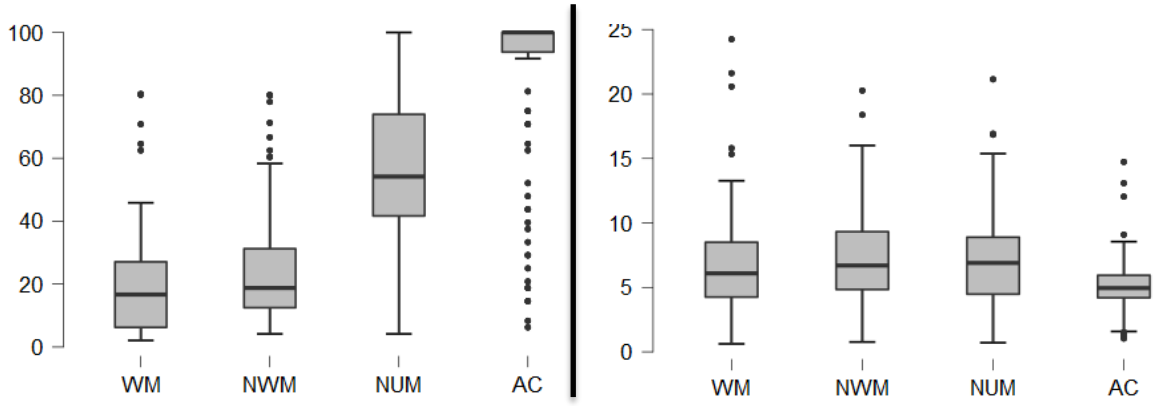
Figures and Tables

Figure 1: Diagram of the multi-group LGCM



Note: Left (Control group; i: intercept; s: linear slope; q: quadratic slope); Right (Training group; G1 and G2 correspond to immediate and long-term treatment effects, respectively)

Figure 2: Boxplots for game progression and gaming time per treatment condition



Left: Percentage of game completed (Y-axis) per treatment condition. Right: Gaming time (in hours) per treatment condition.

Table 1: Model fit indices (top) and parameter estimates of restricted model (and Hedges' g)

	WM-Updt	Math	N-Line (0-10)	N-Line (0-100)	Num Discr
Model fit (AIC/BIC)					
Unrestricted	5608/ 5779	4871/ 5066	11772/ 11967	12486/ 12656	15217/ 15412
Restricted (fixed intercept)	5607/ 5765	4868/ 5051	11768/ 11951	12484/ 12642	15215/ 15398
Means					
Normative Intercept	-.01	-.01	14.17***	18.85***	64.41***
Normative Linear Slope	.89***	1.01***	-2.65***	-1.86**	6.89***
Normative Quadr Slope	-.07***	-.01***	.37***	.01	-1.13***
Variances					
Normative Intercept	.35***	.75***	12.79**	40.98***	126.02***
Normative Linear Slope	.02	.07	3.73*	.89	11.29
Normative Quadr Slope	-[a]	.01	.11*	-[a]	.29
Covariances					
Intercept with L-Slope	.04**	-.05	-2.99	-3.77**	-19.08
% of game completed on G1 and G2					
G1_WM	.05*	.01	-.24**	-.12	-.23
G1_NWM	.03	.04**	-.21**	-.20	.44**
G1_NUM	.03*	.01	-.11*	-.09	.24
G2_WM	.02	.01	.08	-.04	.94
G2_NWM	-.01	-.02	.18	-.21	-.20
G2_NUM	-.01	.01	.11	-.20	.41
Treatment effects					
G1_WM	.02 (.04)	.03 (.04)	.14 (.04)	-.21 (.03)	-.91 (.08)
G1_NWM	-.02 (.03)	-.01 (.01)	-.38 (.11)	-.22 (.03)	-.90 (.08)
G1_NUM	-.02 (.03)	.05 (.06)	-.47* (.13)	-.58* (.09)	-.94 (.08)
G2_WM	-.07 (.13)	-.06 (.07)	-.71 (.20)	-.45 (.07)	1.71 (.15)
G2_NWM	.02 (.04)	.01 (.01)	-.19 (.05)	1.50* (.23)	2.20* (.19)
G2_NUM	.01 (.02)	.01 (.01)	.12 (.03)	1.15* (.18)	2.34 (.21)

Note: *** $p < .001$; ** $p < .01$; * $p < .05$; G1 and G2 correspond to immediate effect (pre- to post-test) and delayed or long-term effect (post-test to delayed post-test), respectively.

[a] Indicates that the variance of the quadratic factor was not estimated due to convergence problems.

Table 2: Conditional effect of the covariate *% of game completed* (and Hedges' g)

	WM-Updt	**Math**	**N- Line (0-10)**	**N-Line (0-100)**	**Num Discr**
G1_WM_30%	.08 (.13)	.04 (.05)	-.10 (.03)	-.33 (.05)	-1.14 (.10)
G1_NWM_30%	.01 (.02)	.03 (.04)	-.59** (.17)	-.42 (.07)	-.47 (.04)
G1_Num_80%	.05 (.08)	.08* (.09)	-.70** (.19)	-.77* (.12)	-.46 (.04)
G2_WM_30%	-.06 (.10)	-.06 (.07)	-.62 (.17)	.41 (.06)	2.65 (.23)
G2_NWM_30%	.01 (.02)	-.01 (.01)	-.01 (.00)	1.30 (.20)	2.01 (.18)
G2_Num_80%	-.01 (.02)	.01 (.01)	.10 (.03)	.75 (.12)	3.16 *(.28)

Note: *** $p < .001$; ** $p < .01$; * $p < .05$; G1 and G2 correspond to immediate effect (pre- to post-test) and delayed or long-term effect (post-test to delayed post-test), respectively.