Supplementary Material

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

## CONSORT flow diagram

## Intervention “Space Ranger Adventures”

Each treatment condition (WM, NWM, NUM, and AC) was conceptualized as a tablet game that incorporated stories and themes to mimic a video-game-like setting. Each game included twelve different phases (or minigames) that varied in terms of game concept or theme or action that was required (e.g., matching objects, collecting and ordering) and types of stimuli that were used (e.g., concrete/abstract objects, symbolic/non-symbolic numerical magnitudes). Within each phase or minigame, difficulty increased as a function of the number of stimuli that had to be recalled or working memory span (in WM and NWM game) and numerical aspects such as number size, the ratio between numbers, or arithmetic operation that was required (in NUM and NWM game). The AC game was a parallel version of the WM game, but the recall component was removed—e.g., children only had to match pairs of objects. Both WM and NWM games were similar in terms of game concept. However, stimuli in the NWM were numerical (similar to those in the NUM training)—e.g., instead of matching objects, the child had to match numbers.

Each game was presented as a space adventure to maintain children’s engagement. Participants played the role of a fresh graduate space ranger tasked to perform different missions. Each mission related to a different phase or minigame (12 missions). Rewards were given after completing each mission (‘currencies’ to unlock upgrades and improve different elements of the game such as spaceship and hangar). Progression through each mission was also rewarded with ‘badges’ to upgrade the status of space ranger.

The aims of the twelve missions were based on four different scenarios/themes or game concepts with different stories and actions (matching objects/numbers, collecting and ordering objects/numbers, connecting objects/numbers, selecting objects/numbers). Differences between missions that were based on the same game concept related to the type of stimulus that was displayed (e.g., matching concrete objects such as fruits, matching abstract objects such as aliens and matching elements that were depicted differently such as a drawing of an apple and a picture of an apple).

In WM and NWM games, each of those game concepts corresponded to a different WM-Updating paradigm (i.e., running span, keep track, *n*-back, and complex span). For instance, *matching* missions required children to recall a specific number of items at the end of a list of sequentially presented stimuli (running span paradigm). *Collecting and ordering* missions required children to store and simultaneously categorize items that were recalled at a later stage (keep track paradigm). *Connecting* missions required children to indicate whether the current stimulus was the same as the one which appeared n items earlier (*n*-back paradigm) and *selecting* missions required children to make decisions on the nature of a pair of items (e.g., edible, larger) as well as storing spatial information regarding the location of the targets (complex span paradigm). Across all four types of missions, in WM and NWM games, children had to continuously update the items that had to be recalled at a later stage (1 to 4, corresponding to span 1 to span 4, respectively) since they were presented as part of a list of items (or pairs of items in complex span) of unknown length.

In the Numeracy (NUM) and Active Control (AC) treatment conditions, the missions did not include recalling but required similar actions. For instance, children had to match symbolic and non-symbolic numbers (NUM game), or objects/pictures such as those presented in the WM game (AC game).

**Game concept**

***Matching* *missions***

The *matching* missions related to a space villain who had kidnapped innocent aliens. The space ranger was tasked to go into the villains’ cave and rescue the aliens. Across games or treatment conditions, the space ranger had to unlock a series of gates by identifying/matching the correct key card that was depicted on top of each gate, from a set of three key cards. In the WM and NWM training children were tasked to remember (and recall) the keys that unlocked, for instance, the last two gates (span 2). Figure S-1 is a screenshot that shows an example of how children were prompted to recall in the mission of the NWM game that included symbolic numerical stimuli. In the NUM training, the space ranger was required to select the key card that represented the magnitude of the dot array or number depicted on each gate (or vice versa). That is, transcoding between symbolic and non-symbolic numerical representations. Children undergoing AC training only matched similar items as in the WM training game. No recall was required.

*Figure S-1:* NWM game (running span paradigm)



***Collecting and ordering* *missions***

*Collecting and ordering* missions required the space ranger to defeat a giant monster that was destroying Space City. The monster was throwing objects to the space ranger. Children were tasked to collect the objects and get them stored in their corresponding locations at the bottom of the screen (e.g., pears with pears, apples with apples). In the WM and NWM training games, these items were shown encapsulated into different shapes (e.g., triangle, square) that served as categories, and children were required to remember (and recall), for instance, the last item that was shown in a circle and the last one that was shown in a square (span 2). The number of categories (shapes) increased as a function of items to be recalled (i.e., only 2 shapes served to encapsulate the items in span 2). Figure S-2 is a screenshot that shows an example of how these categories were created in the WM training game that included concrete objects (e.g., fruits). The NUM training (and NWM training game) involved a number line. Children were required to identify the correct position of the number that was falling. Children undergoing the AC training only collected and ordered the items that were falling as in the WM training game. No recall was required.

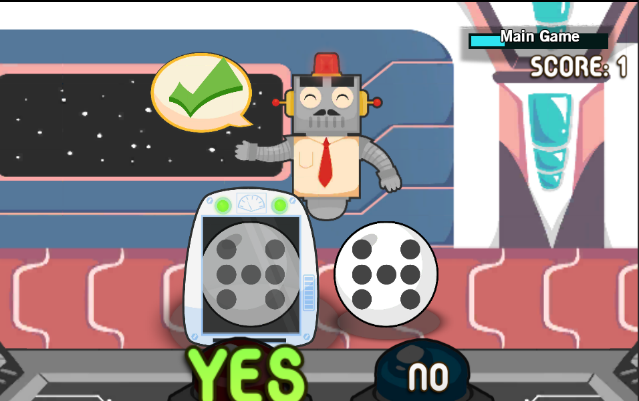
*Figure S-2:* WM game (keep track paradigm)



***Connecting* *missions***

*Connecting* missions required the space ranger to help out at Planetary Mailroom by deciding if similar items had been shipped out previously. Children were shown a conveyor belt that showed the content of one box at a time. Children undergoing WM and NWM training had to check whether the item that was presented in a box was similar to that displayed, for instance, two boxes before (span 2). Figure S-3 is a screenshot that shows how feedback was conceptualized in the NWM training game that included non-symbolic stimuli and required children to connect with the previously shown stimulus (1-back)—the target is unveiled in the feedback screen. In the NUM training, the content of the boxes was always visible and children were tasked to decide whether the quantity displayed on the following box corresponded to a specific counting sequence (e.g., counting by twos). The AC training stimuli followed the WM training stimuli, but the content of all boxes always remained visible.

*Figure S-3:* NWM game (n-back paradigm)



***Selecting* *missions***

*Selecting* missions required the space ranger to select edible items in the WM and AC games and the larger of two quantities in the NWM and NUM games. In the WM and NWM games, each pair of items/numbers were randomly arranged on a 3x3 grid (see Figure S-4, left panel) and children were tasked to select the edible pair (or larger number in the NWM game), and memorize the location for later recall. In the NUM and AC games, no grid was presented. (see Figure S-4, right panel). In the NUM training, children were shown with two pairs of numerical stimuli (e.g., Arabic numbers, equations, arrays of dots) and asked to select the larger in terms of numerical magnitude.

*Figure S-4:* Selecting mission in WM (top row), NWM (second row), NUM (third row) and AC game (bottom row)A close-up of a circuit board

Description automatically generated with medium confidence

*Note:* “…” indicates that several trials were presented before the recall phase. The third row shows instances of both symbolic and non-symbolic stimuli.

**Game Progression**

Tutorials and practice trials were inserted at the beginning of the game and before each phase/mission and difficulty level. Children were not shown actual trials until successfully passing two out of three practice trials. Game progression was auto-adjusted to children’s performance and pre-configured to provide children with experience in all twelve types of missions (four game concepts x three types of stimuli).

In WM and NWM games, four different levels of difficulty were presented in each mission/phase, corresponding to span 1 to 4. For instance, children playing the mission that required matching symbolic numbers and their corresponding non-symbolic representations (NWM game) were challenged with four different levels of increasing difficulty that corresponded to recalling from 1 to 4 items (i.e., span 1 to 4). Each level of difficulty within that particular mission was passed after successfully recalling the corresponding number of items in 4 out of the last 6 lists of items that were presented. After completing span 4, children were rewarded and allowed to upgrade or improve some elements of the game. Then, they were presented with the tutorial and practice trial of a different mission at the entry level of difficulty (e.g., selecting and ordering symbol numbers on a number line and remembering the last number that was presented in a triangle). When children struggled with a particular difficulty level (i.e., failing to recall correctly the last two abstract items in a list for six consecutive times), they were presented with a less challenging difficulty level (recalling 1 item) and a different mission that involved easy-to-remember items (e.g., fruits). The difficulty level in the NUM training (within each phase and mission) was a function of number size, the ratio between numbers and speed of presentation.

## Descriptive statistics

All descriptive statistics were estimated using Mplus (Version 8.3; Muthén & Muthén, 1998–2017). Table S-1 shows the descriptive statistics (per treatment conditions and time point) for all of the WM-Updating measures, math measures, as well as performance measures on the Number-Line task and Numerical Discrimination task. Test-retest reliability values (ICC) for all of the outcome variables were “moderate” to “good” (aggregated across treatment conditions).

*Table S-1:* Descriptive statistics (per treatment condition and time point)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **WM** | | | | **NWM** | | | | **NUM** | | | | **AC** | | | |  |
|  | **Time point** | **Min Max** | | **M** | **SD** | **Min Max** | | **M** | **SD** | **Min Max** | | **M** | **SD** | **Min Max** | | **M** | **SD** | **ICCa**  **(95% CI)** |
| ***WM outcomes* (raw scores)**  Running Letters | 1 | .0 | 21.0 | 9.6 | 3.9 | .0 | 16.0 | 9.3 | 3.4 | .0 | 18.0 | 9.1 | 4.2 | .0 | 18.0 | 9.5 | 3.7 | .66  (.47, .77) |
|  | 2 | 1.0 | 22.0 | 11.8 | 4.3 | 1.0 | 19.0 | 11.2 | 4.1 | 1.0 | 18.0 | 10.7 | 3.8 | 1.0 | 23.0 | 11.1 | 4.3 |
|  | 3 | 1.0 | 23.0 | 13.9 | 4.1 | 1.0 | 23.0 | 13.6 | 4.2 | 4.0 | 22.0 | 13.4 | 3.8 | 6.0 | 21.0 | 14.1 | 3.6 |  |
|  | 4 | .0 | 23.0 | 14.4 | 4.4 | 5.0 | 22.0 | 14.7 | 4.0 | 6.0 | 21.0 | 14.0 | 3.7 | 2.0 | 23.0 | 14.5 | 4.2 |  |
| Keep Track | 1 | .0 | 17.0 | 5.2 | 2.8 | .0 | 11.0 | 5.0 | 2.4 | .0 | 13.0 | 4.9 | 2.5 | .0 | 12.0 | 5.7 | 2.5 | .69  (.48, .80) |
|  | 2 | .0 | 17.0 | 7.1 | 3.4 | 1.0 | 13.0 | 6.7 | 3.0 | .0 | 16.0 | 6.1 | 2.8 | .0 | 17.0 | 7.2 | 3.1 |
|  | 3 | .0 | 17.0 | 8.7 | 3.3 | 1.0 | 17.0 | 8.4 | 3.3 | .0 | 14.0 | 7.8 | 2.9 | 1.0 | 17.0 | 8.8 | 3.2 |  |
|  | 4 | 3.0 | 19.0 | 9.7 | 3.1 | 2.0 | 17.0 | 9.3 | 3.1 | .0 | 19.0 | 8.8 | 3.4 | .0 | 17.0 | 9.3 | 3.3 |  |
| N-Back | 1 | .0 | 41.0 | 22.1 | 8.6 | 1.0 | 42.0 | 21.5 | 7.6 | 5.0 | 42.0 | 22.6 | 7.9 | 7.0 | 44.0 | 23.3 | 7.9 | .68  (.60, .75) |
|  | 2 | 11.0 | 44.0 | 25.7 | 6.6 | 7.0 | 42.0 | 25.3 | 6.8 | 6.0 | 43.0 | 25.5 | 6.7 | 6.0 | 39.0 | 24.8 | 6.7 |
|  | 3 | 7.0 | 42.0 | 26.0 | 6.5 | 10.0 | 41.0 | 25.8 | 6.8 | 8.0 | 39.0 | 26.9 | 6.4 | 10.0 | 42.0 | 27.1 | 6.7 |  |
|  | 4 | 11.0 | 40.0 | 27.3 | 5.9 | 7.0 | 39.0 | 28.0 | 6.3 | 5.0 | 43.0 | 28.4 | 6.8 | 7.0 | 41.0 | 28.1 | 7.0 |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Complex Span | 1 | .0 | 21.0 | 4.8 | 3.8 | .0 | 17.0 | 5.0 | 4.0 | .0 | 24.0 | 4.8 | 3.6 | 1.0 | 17.0 | 5.7 | 4.0 | .65  (.43, .77) |
|  | 2 | .0 | 22.0 | 6.8 | 5.0 | .0 | 24.0 | 7.3 | 5.2 | .0 | 22.0 | 6.9 | 5.3 | .0 | 21.0 | 7.3 | 5.1 |
|  | 3 | .0 | 23.0 | 10.5 | 5.9 | .0 | 24.0 | 10.4 | 6.3 | .0 | 24.0 | 9.8 | 6.0 | .0 | 24.0 | 10.9 | 6.0 |  |
|  | 4 | 1.0 | 24.0 | 12.2 | 6.0 | 1.0 | 24.0 | 12.4 | 6.2 | 1.0 | 24.0 | 12.3 | 6.4 | 1.0 | 24.0 | 13.7 | 5.7 |  |
| ***Early numeracy***  N-Line 0-10 (PAE) | 1 | 3.4 | 36.1 | 14.6 | 6.4 | 2.5 | 37.4 | 14.8 | 7.1 | 4.8 | 28.9 | 13.8 | 5.4 | 2.8 | 35.6 | 13.4 | 6.2 | .62  (.48, .71) |
|  | 2 | 3.1 | 24.5 | 11.9 | 4.8 | 2.3 | 29.7 | 11.2 | 5.4 | 2.4 | 22.6 | 11.4 | 4.8 | 2.4 | 25.7 | 11.3 | 5.2 |
|  | 3 | 3.3 | 31.9 | 10.2 | 5.2 | 1.6 | 21.8 | 8.5 | 4.4 | 2.0 | 26.2 | 8.5 | 4.2 | 2.1 | 26.8 | 9.4 | 4.9 |  |
|  | 4 | 2.0 | 21.5 | 9.1 | 4.7 | 1.8 | 18.4 | 8.0 | 4.1 | 2.0 | 19.5 | 8.1 | 3.9 | 2.2 | 21.5 | 9.3 | 5.0 |  |
| N-Line 0-100 (PAE) | 1 | 4.0 | 40.4 | 19.1 | 8.4 | 5.5 | 37.2 | 20.5 | 7.9 | 5.6 | 38.4 | 18.8 | 7.4 | 4.6 | 43.5 | 18.9 | 8.2 | .79  (.65, .86) |
|  | 2 | 4.2 | 39.8 | 16.5 | 7.8 | 3.8 | 35.3 | 16.3 | 7.4 | 5.3 | 34.5 | 15.6 | 7.4 | 2.7 | 42.5 | 15.6 | 8.7 |
|  | 3 | 4.3 | 32.7 | 13.3 | 6.5 | 2.7 | 36.2 | 13.5 | 7.5 | 3.8 | 27.5 | 12.4 | 5.6 | 2.4 | 35.0 | 12.7 | 6.8 |  |
|  | 4 | 3.9 | 30.8 | 11.5 | 5.7 | 3.3 | 35.5 | 13.0 | 6.9 | 3.3 | 32.6 | 11.5 | 5.6 | 3.5 | 45.9 | 11.6 | 6.6 |  |
| Num Discr (% accuracy) | 1 | 26.0 | 93.0 | 63.6 | 15.9 | 26.0 | 96.0 | 63.1 | 16.4 | 31.0 | 94.0 | 66.2 | 14.8 | 12.0 | 95.0 | 64.6 | 16.0 | .73  (.66, .78) |
|  | 2 | 17.0 | 94.0 | 68.9 | 15.8 | 18.0 | 93.0 | 68.5 | 15.5 | 41.0 | 95.0 | 71.7 | 14.1 | 36.0 | 94.0 | 71.9 | 13.7 |
|  | 3 | 12.0 | 94.0 | 72.0 | 14.3 | 22.0 | 92.0 | 72.4 | 13.6 | 12.0 | 94.0 | 72.7 | 15.8 | 31.0 | 95.0 | 74.5 | 14.1 |  |
|  | 4 | 24.0 | 94.0 | 73.5 | 13.9 | 33.0 | 96.0 | 74.5 | 12.5 | 37.0 | 96.0 | 75.6 | 13.4 | 17.0 | 95.0 | 74.2 | 15.6 |  |
| ***Math outcomes* (raw scores)**  Math Problem Solving | 1 | 15.0 | 35.0 | 25.2 | 4.7 | 14.0 | 35.0 | 26.1 | 4.8 | 15.0 | 37.0 | 26.1 | 4.8 | 15.0 | 36.0 | 26.6 | 4.3 | .64  (.21, .81) |
|  | 2 | 18.0 | 40.0 | 29.7 | 4.3 | 20.0 | 39.0 | 30.1 | 3.9 | 20.0 | 42.0 | 30.8 | 4.3 | 19.0 | 40.0 | 30.3 | 4.4 |
|  | 3 | 22.0 | 46.0 | 33.9 | 4.7 | 22.0 | 46.0 | 34.0 | 3.9 | 26.0 | 46.0 | 34.3 | 3.9 | 19.0 | 45.0 | 34.3 | 4.8 |  |
|  | 4 | 21.0 | 45.0 | 35.2 | 4.3 | 23.0 | 46.0 | 35.5 | 4.3 | 26.0 | 46.0 | 35.8 | 4.0 | 23.0 | 49.0 | 35.9 | 4.5 |  |
| Numerical Operations | 1 | 3.0 | 23.0 | 13.7 | 3.8 | 3.0 | 23.0 | 13.9 | 3.4 | 2.0 | 22.0 | 13.8 | 3.5 | 2.0 | 24.0 | 14.3 | 4.0 | .71  (.28, .86) |
|  | 2 | 7.0 | 28.0 | 17.3 | 4.1 | 2.0 | 25.0 | 17.1 | 3.6 | 5.0 | 26.0 | 17.3 | 3.9 | 5.0 | 26.0 | 17.5 | 4.0 |
|  | 3 | 10.0 | 29.0 | 20.7 | 4.3 | 8.0 | 31.0 | 20.6 | 4.4 | 9.0 | 30.0 | 20.9 | 3.8 | 4.0 | 31.0 | 21.3 | 4.4 |  |
|  | 4 | 4.0 | 32.0 | 21.7 | 4.6 | 10.0 | 32.0 | 21.7 | 4.3 | 14.0 | 32.0 | 22.5 | 4.1 | 4.0 | 30.0 | 22.3 | 4.8 |  |
| Math Fluency | 1 | .0 | 58.0 | 14.0 | 10.9 | .0 | 47.0 | 13.6 | 10.2 | .0 | 46.0 | 13.6 | 9.4 | 1.0 | 49.0 | 15.8 | 11.4 | .84  (.73, .92) |
|  | 2 | .0 | 62.0 | 19.6 | 12.8 | 3.0 | 58.0 | 21.2 | 12.2 | .0 | 49.0 | 20.5 | 12.6 | 1.0 | 55.0 | 22.0 | 12.6 |
|  | 3 | .0 | 57.0 | 26.4 | 12.6 | 1.0 | 68.0 | 27.0 | 13.8 | 4.0 | 62.0 | 27.2 | 12.3 | 1.0 | 55.0 | 28.5 | 13.4 |  |
|  | 4 | 1.0 | 61.0 | 31.3 | 13.3 | 5.0 | 69.0 | 33.4 | 13.7 | 7.0 | 65.0 | 32.9 | 13.5 | 2.0 | 62.0 | 32.9 | 13.6 |  |

Note: PAE refers to percentage of absolute error (higher values indicate worse performance); *d prime* (*d’*) is a sensitivity index that reflects the ability to discriminate both correct and false responses (higher values indicate better performance). a Single-measurement, absolute-agreement, 2-way mixed-effects model.

## Preliminary analyses

All descriptive and inferential statistical analyses were estimated using Mplus (Version 8.6; Muthén & Muthén, 1998–2017).

### **WM-Updating Measurement Invariance (over time)**

We tested three levels of measurement invariance that corresponded to increasingly restrictive hypotheses. Configural invariance involves testing whether the same item-factor structure (or CFA model) is applicable across groups (over time). Metric invariance builds upon configural invariance and requires testing whether the factor loadings of those items are also equivalent across groups (i.e., constraining the factor loading of each item to equality across groups). Scalar invariance builds upon metric invariance by requiring that the item intercepts also be equivalent across groups. Attaining scalar invariance is a pre-requisite to explore differences over time at the latent factor mean levels.

Goodness of fit was assessed by inspecting the *χ2* test, as well as the Comparative Fit Index (CFI) and the Root Mean Square Error of Approximation (RMSEA). For CFI, we assumed values above 0.95 to indicate adequate fit (Hu & Bentler, 1999). For the RMSEA, we considered values below 0.06 as indicating a good model fit (Chen et al., 2008). In evaluating the measurement invariance assumptions, we also relied on Cheung and Rensvold’s (2002) criterion (if the change in CFI is less than .01, then invariance holds).

The configural model fitted the data well (see Table S-2). A pair of residual variances were allowed to covary at each time point (*N*-Back and Complex Span). Nonetheless, the more restrictive models to test metric and scalar invariance did not provide an acceptable fit to the data. Furthermore, the decrease in CFI and RMSEA was larger than that suggested above.

*Table S-2*: Model fit for the invariance testing

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **χ2 (df)** | **RMSEA (Prob** < **.05)** | **CFI** |
| Configural | 7.51(4)ns | .04 (.48) | .996 |
| Metric | 39.53 (13)\* | .07 (.09) | .969 |
| Scalar | 113.59 (22)\* | .10 (.00) | .893 |
| Metric vs. Configural | 32.85 (9)\* | - | - |
| Scalar vs. Configural | 108.90 (18)\* | - | - |
| Scalar vs. Metric | 77.39 (9) \* | - | - |

*Note:* ns. Denotes non-significant. \* <. 001

Next, we investigated whether partial measurement invariance was tenable. Given that complete scalar invariance is sometimes an unachievable goal, it has been argued that constructs that are approximately or partially invariant across groups or over time may be deemed as reliable (Byrne, Shavelson, & Muthén, 1989). Rather than looking at partial or approximate invariance with approaches that are based on stepwise procedures that have drawn severe criticisms due to their caveats (see Byrne et al., 1989), we used the alignment method (FIXED option) developed by Asparouhov and Muthén, (2014) to determining partially invariant measurement parameters, the possible set of groups in which invariance holds, and whether both latent variables can be considered reliable (i.e., whether we are measuring the same construct over time). The aligned model has the same fit as the configural model. In their study, Asparouhov and Muthén found that, point estimates were unbiased for large group sizes (1000), and that the biases for small group sizes (100) were not large and tended to occur only when the amount of non-invariance was large. For a small number of groups (time points) and few non-invariances (less than 20-25%; Asparouhov & Muthén, 2014), the results of the alignment method are trustworthy.

The alignment revealed that only one loading parameter remained significantly non-invariant after the alignment of the WM-Updating construct, which is 3% of the total number of parameters that were estimated (and well below the 20-25% cut off). The average invariance index was .79 (1 stands for perfect scalar invariance), which suggests appropriate partial invariance over time, or that the factor means that are estimated afford a trustworthy comparison over time. Table S-3 shows the invariance results for aligned intercept and loading parameters.

*Table S-3*: Approximate Measurement (Non) Invariance for Intercepts and Loadings Over Time and Alignment fit statistics

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | **Intercept** | | | | | **Loading** | | | |
|  | | **Time Point** | | ***R2*** |  | | **Time Point** | | | ***R2*** | |
| Letter Running | 1 2 3 4 | | .99 | | |  | | 1 2 3 4 | .28 | | | |
| Keep Track | 1 2 3 4 | | 1.00 | | |  | | 1 2 3 4 | .97 | | | |
| N-Back | 1 2 3 4 | | .89 | | |  | | 1 2 3 4 | .87 | | | |
| Complex Span | 1 2 3 4 | | .86 | | |  | | (1) 2 3 4 | .47 | | | |

*Note: R*2 indicates the degree of invariance for a parameter. A high value indicates a high degree of measurement invariance (Asparouhov & Muthén, 2014). Bracket indicates invariance.

### **Math Measurement Invariance (over time)**

We proceed in a similar way to test the measurement invariance of the math construct that was formed with the three outcomes of the WIAT. Given that only three indicators were considered, the configural model did not provide fit information. Inspection of the amount of explained variance for each indicator (and across waves) showed values above 50% (standardized loadings above .70), which indicates that the latent construct reflects quite well the underlying causes of the three observed variables. The model to test metric invariance showed an acceptable fit to the data (χ2 (6) = 15.78, CFI = .995, TLI = .990, RMSEA = .062 [.025, .101]). Nonetheless, the most restrictive models to test scalar invariance did not provide an acceptable fit (χ2 (12) = 190.36, CFI = .905, TLI = .905, RMSEA = .188 [.165, .212]). Furthermore, the change in CFI was well above the criterion suggested by Cheung and Rensvold (2002).

The alignment method revealed 3 non-invariant parameters (12.5%; see Table S-4). The average invariance index was .79, which suggests appropriate partial invariance over time.

*Table S-4*: Approximate Measurement (Non) Invariance for Intercepts and Loadings Over Time and Alignment fit statistics

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Intercept** | | | | | | **Loading** | | | | |
|  | **Time Point** | | ***R2*** | |  | | **Time Point** | | | ***R2*** | | |
| Numerical Operations | | 1 2 (3) 4 | | .99 | |  | | 1 2 3 4 | .92 | |
| Math Problem Solving | | (1) 2 3 4 | | .95 | |  | | 1 2 3 4 | .62 | |
| Math Fluency | | (1) 2 3 4 | | .84 | |  | | 1 2 3 4 | .46 | |

*Note: R*2 indicates the degree of invariance for a parameter. A high value indicates a high degree of measurement invariance (Asparouhov & Muthén, 2014). Brackets indicate invariance.

### **Single-group latent growth curve model**

We estimated four single-group latent growth curve models for each outcome variable to investigate whether similar longitudinal trends existed across groups. The TSCORES option in Mplus was used to account for time-varying observations. This method only provided relative fit indices (the Akaike information criterion, AIC, Akaike, 1973; and the Bayesian information criterion, BIC, Schwarz, 1978). Differences in AIC and BIC values afford informal comparisons of models with differing numbers of parameters. Smaller BIC (Δ > 2) and AIC (Δ > 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). In the context of modeling single-group data, this relates to investigating whether a model with a linear trend fits the data better than a model with a quadratic term. Smaller BIC suggests a better model fit. Table S-5 shows the relative model fit statistics.

*Table S-5:* Model fit statistics (AIC/BIC) for the single-group models

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **WM-Updating** | **Math** | **N-Line**  **(0-10)** | **N-Line (0-100)** | **Numeric Discrimin** |
| WMa | 990/1014 | 877/901 | 2501/2525 | 2600/2624 | 3353/3377 |
| WMb | 958/993 | 789/824 | 2498/2533 | 2596/2631 | 3348/3383 |
| NWMa | 977/1001 | 810/834 | 2484/2508 | 2706/2701 | 3290/3314 |
| NWMb | 953/988 | 738/773 | 2463/2498 | 2692/2727 | 3289/3324 |
| NUMa | 930/954 | 795/819 | 2404/2429 | 2592/2617 | 3287/3311 |
| NUMb | 901/936 | 695/730 | 2399/2434 | 2578/2613 | 3291/3326 |
| ACa | 953/977 | 881/905 | 2541/2565 | 2740/2764 | 3410/3434 |
| ACb | 940/975 | 786/821 | 2540/2574 | 2731/2766 | 3403/3437 |

*Note:* a and b denote linear and quadratic trends, respectively.

For all of the outcome variables, and across treatments, the means of the linear and quadratic slopes were significantly different from zero. The variance of the quadratic term was only significant in the Number-Line (0-100) estimation task (all but the NWM condition). Independently of outcome measure, the AC group showed a deceleration in the growth rate over time.

**References**

Akaike, H. (1973). *Information theory and an extension of the maximum likelihood principle*. In B. N. Petrov & F. Caski (Eds.), Proceedings of the Second International Symposium on Information Theory (pp. 267-281). Budapest: Akademiai Kiado.

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(4), 495-508.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456.

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research, 36*(4), 462-494.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural equation modeling, 9*(2), 233-255.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: a Multidisciplinary Journal, 6*(1), 1-55.

Muthén, B., & Asparouhov, T. (2018). Recent methods for the study of measurement invariance with many groups: Alignment and random effects. *Sociological Methods & Research, 47*(4), 637-664.

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

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, 111-163.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2),461-464.

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.