Article type : Paper (under 8000 words)

Predictors of mathematics in primary school: magnitude comparison, verbal and spatial working memory measures

Sara Caviola [1,2]*, Lincoln J. Colling [2,3], Irene C. Mammarella [4], Dénes Szűcs [2]*

[1] *School of Psychology, University of Leeds, UK*

[2] *Centre for Neuroscience in Education, Department of Psychology, University of Cambridge, UK*

[3] *School of Psychology, University of Sussex, UK*

[4] *Department of Developmental Psychology, University of Padova, Italy*

*Correspondence to:

(1) Dénes Szucs: ds377@cam.ac.uk

Centre for Neuroscience in Education,

Department of Psychology, Downing Street,

Cambridge CB2 3EB, UK

(2) Sara Caviola: s.caviola@leeds.ac.uk

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/DESC.12957

School of Psychology
University of Leeds
Leeds, LS2 9JT, UK

Research Highlights (max 4 - 25 words each)

- In a large cross-sectional study (≃1200 children) we determined relationships between magnitude comparison, working memory capacity, standardized math and reading achievement.
- We provide evidence for the lack of association between non-symbolic magnitude comparison measures and mathematics achievement.
- Symbolic number comparison accuracy and spatial working memory were specifically associated with mathematical performance.
- Verbal short-term and working memory were associated with both reading and math performance.

Abstract (200/max 250 words)

We determined the relative importance of the so-called approximate number system, symbolic number comparison and verbal and spatial short-term and working memory capacity for mathematics achievement in 1254 Grade 2, 4 and 6 children. The large sample size assured high power and low false report probability and allowed us to determine effect sizes precisely. We used reading decoding as a control outcome measure to test whether findings were specific to mathematics. Bayesian analysis allowed us to provide support for both null and alternative hypotheses. We found very weak zero-order correlations between approximate number system (ANS) measures and math achievement. These correlations were not specific to mathematics, became non-significant once intelligence was considered, and ANS measures were not selected as predictors of math by regression models. In contrast, overall symbolic number comparison accuracy and spatial working memory measures were reliable and mostly specific predictors of math achievement. Verbal short-term and working memory and symbolic number comparison reaction time were predictors of both reading and math achievement. We conclude that ANS tasks are not suitable as measures of math development in school-age populations. In contrast, all other cognitive functions we studied are promising markers of mathematics development.

Keywords: magnitude comparison, mathematics, working memory, correlation, children, Bayesian.

Identifying strong correlates and potential predictors of mathematical development has important theoretical and practical relevance. A proposed domain specific predictor that has received a lot of attention during the past 20-30 years is the so-called evolutionarily based 'number sense' (Dehaene, 1997; Leibovich, Katzin, Harel, & Henik, 2017), a non-symbolic magnitude representation or approximate number system (hereafter ANS). Another proposed domain specific predictor of mathematical development is symbolic number comparison (SNC) ability (Ansari, 2008). Regarding domain general factors, verbal and spatial short-term memory (STM) and working memory (WM) are often thought to be some of the most reliable correlates of mathematical skill (Barrouillet 2018; Caviola, Mammarella, Lucangeli, & Cornoldi, 2014; Cragg & Gilmore, 2014; Fias & Menon, 2013; Menon, 2016; Friso-van den Bos, Van Der Ven, Kroesbergen, & Van Luit, 2013; Passolunghi, Mammarella, & Altoè, 2008; Peng, Namkung, Barnes, & Sun, 2016; Raghubar, Barnes & Hecht, 2010; Szűcs, Devine, Soltesz, Nobes, & Gabriel, 2014). To date, there is no agreement on the relative importance of the above predictors for mathematical development. Most importantly, many question whether the ANS plays any specific role in school relevant mathematics while others argue that it is one of the most important factors (Feigenson, Dehaene, & Spelke, 2004; Halberda, Ly, Wilmer, Naiman, & Germine, 2012; Halberda, Mazzocco, & Feigenson, 2008; Hohol, Cipora, Willmes, & Nuerk, 2017; Leibovich et al., 2017; Szűcs et al., 2014). Progress is precluded by the fact that the field lacks large studies with high statistical power and precise effect size estimates that test not only the ANS but also other relevant cognitive factors. Here we report such a study of 1254 Grade 2, 4 and 6 children. We have carried out both null hypothesis significance testing and Bayesian analysis, the latter being able to quantify support for both the null and alternative hypotheses. Our study represents a strong test of recent theoretical and empirical models that have included domain specific (ANS and SNC) and domain general factors (WM) as predictors of mathematics achievement (Geary, 2013; Goffin & Ansari, 2019; Inglis, Attridge, Batchelor & Gilmore, 2011).

The ANS is often investigated in magnitude comparison tasks (Lyons, Nuerk, & Ansari, 2015; Price, Palmer, Battista, & Ansari, 2012; Smets, Sasanguie, Szücs, & Reynvoet, 2015). In a typical task, participants decide which of two visually presented groups of items is more numerous. The most frequent measures of ANS are the proportion of correct numerical decisions (accuracy) and the so-called Weber-fraction (*w*), a measure derived from accuracy (lower accuracy corresponds to higher *w*). *w* characterizes the shape of a model based sigmoid curve fitted to accuracy data. This sigmoid shape depends on the sharpness of discrimination ability.

Importantly, many ANS studies have not considered the influence of confounding visual display parameters when determining *w* (e.g., convex hull, density, size; Gebuis & Reynvoet, 2012; Leibovich & Henik, 2013; Szűcs, Nobes, Devine, Gabriel, & Gebuis, 2013b). This can be done by taking into account which visual parameters are congruent and which ones are incongruent with numerical information (see more details in Methods; for technical reviews see De Smedt, Noël, Gilmore, & Ansari, 2013; Fabbri, et al., 2012; Gebuis & Reynvoet, 2012b; Szűcs et al., 2013b; Tokita & Ishiguchi, 2013). Generally, accuracy and therefore *w* is influenced by the level of congruity between numerosity and continuous visual parameters. This influence is larger in children than in adults (Szűcs et al., 2013b). Notably, recent papers have challenged the validity and reliability of several ANS measures also suggesting that inconsistent findings may be explained by differences in measures (Dietrich, Huber, & Nuerk, 2015; Inglis, Gilmore, 2014; Leibovich et al., 2017; Szűcs et al., 2013b). Hence, here we use several measures of the ANS so that findings are comparable with most of the literature.

SNC ability is typically measured in tasks where participants decide which of two symbolically presented numbers is numerically larger (e.g. which one is larger, 3 or 6?), or whether a number is smaller or larger than a reference number (e.g. is 3 smaller or larger than 5?). It is important to clearly distinguish between various measures that can be derived from SNC tasks. First, some have proposed the use of so-called numerical distance effects (closer numbers [e.g. 6 vs. 5] are slower and more error prone to discriminate than further away numbers [e.g. 9 vs. 5]) that are often thought to be the consequence of the involvement of the ANS in symbolic numerical decisions (Dehaene, 1997; Moyer & Landauer, 1967). Second, many investigators use overall accuracy and reaction time (RT) rather than distance effects. These measures may not necessarily reflect number representation related processes but may rather characterize the accuracy and speed of access to symbolic numerical (and non-numerical) information.

Several studies have investigated whether individual differences in ANS performance or SNC are associated with mathematics development in primary school children, mainly considering one (Xenidou-Dervou, De Smedt, van der Schoot, & van Lieshout, 2013) or two (Gimbert, Camos, Gentaz & Mazens, 2019) age groups or employing longitudinal designs (Wong, Ho & Tang, 2016; Xenidou-Dervou, Molenaar, Ansari, van der Schoot, & van Lieshout, 2017). Current evidence is inconsistent, delivering both positive and negative results and generally small effect sizes (De Smedt et al., 2013; Fias & Menon, 2013; Halberda et al., 2012, 2008; Hohol et al., 2017; Holloway & Ansari, 2008; Leibovich et al., 2017; Menon, 2016; Nosworthy, Bugden, Archibald,

Evans, & Ansari, 2013; Sasanguie, De Smedt, Defever, & Reynvoet, 2012; Sasanguie, de Smedt, & Reynvoet, 2015; Sasanguie, Defever, Maertens, & Reynvoet, 2014; Szűcs et al., 2014).

The mixed nature of results is well exemplified by the outcomes of four recent meta-analyses. Two studies considered only the association between math and the ANS. Both reported weak correlations that varied depending on the types of measures and age groups (r = 0.24 and 0.22 respectively; Chen & Li, 2014; Fazio, Bailey, Thompson, & Siegler, 2014). Schwenk et al., (2017) considered studies of children with and without mathematical difficulties and concluded that only SNC reaction time (RT) but not accuracy discriminated between children with and without math difficulties. Schneider et al., (2017) considered 195 results for ANS and 89 results for SNC tasks. They showed stronger association between math achievement and SNC tasks (r = .302) than with ANS tasks (r = .241, Schneider et al., 2016). There was a lot of heterogeneity between ANS studies (rs ranging between -0.2 – 0.8) and the strength of the association appeared to decrease with age (see Figure 2 in Schneider et al., 2016). Using this data (kindly provided by M. Schneider) we determined that median sample sizes were 64 and 49 in non-symbolic and symbolic studies, respectively, and also computed power distributions for all studies (using Matlab's 'sampsizepwr' function; The MathWorks Inc., Natick MA). We found that only 26% (ANS) and 25% (SNC) of studies were powered at the 0.8 level to detect the effects found in the meta-analysis (α=0.05; one-tailed). For studies employing ANS tasks, median power was 0.625 to detect r = 0.241. For studies employing SNC tasks median power was 0.705 to detect r = 0.302. On their own, these power levels are suboptimal. Additionally, considering that the overwhelming bias for publishing statistically significant results acts as a filter favoring exaggerated effect sizes, published meta-analyses likely overestimate the true magnitude of any relationships (Ioannidis, 2008, 2010; Szűcs & Ioannidis, 2017a). Hence, high powered primary studies are clearly needed to estimate effect sizes precisely.

Overall, data suggest that ANS related findings are highly variable. Researchers often explain variable findings by invoking variability in 'moderator variables', such as age, tasks and math assessment tools. For example, age likely contributes to variability (Dietrich, et al., 2015; Inglis & Gilmore, 2014; for discussion see Schneider et al. 2016). However, the low statistical power of many studies can also explain variability (Tosto et al., 2017). In general, the lower is statistical power the more diverse findings can be expected a priori due to imprecise effect size meaurement. In addition, low power also increases 'false report probability', the chance that statistically significant findings are in fact false (Szűcs & Ioannidis, 2017b). Low statistical power

also seriously limits meta-analyses because these overwhelmingly rely on highly exaggerated and imprecise (noise-prone) published effect sizes measured in underpowered studies (Ioannidis, 2005, 2008; Szűcs & Ioannidis, 2017a, 2017b). Therefore, high powered studies are necessary to provide precise and reliable magnitude and interval estimates for effects. Additionally, as most psychological constructs are likely to be correlated, it is more meaningful to contrast the relationships of alternative constructs than only measure one relationship only (see Meehl, 1967; Szűcs & Ioannidis, 2017b). This is very important to consider when interpreting meta-analyses that are usually unable to consider multiple measures due to the use of diverse measurement constructs in studies.

Considering multiple variables simultaneously is especially important in a complex domain such as mathematical development that likely relies on an extended network of cognitive skills (Fias, Menon, & Szűcs, 2013; Krajewski & Schneider, 2009; Szűcs et al., 2014; Xenidou-Dervou et al., 2018). A number of studies suggest that the relationship between ANS and math may be explained by executive functions (inhibitory and attentional control and WM) contributing to both math and ANS (Fuhs & McNeil, 2013; Gilmore et al., 2013; Geary, 2013; Price & Wilkey, 2017). An option is that ANS may support symbolic numerals' acquisition of meaning by mapping them onto analogue magnitudes (Geary, 2013). Attentional control may become more relevant after practice with SNC and other math domains. Similarly, Inglis et al., (2011) suggested that the relationship between ANS and math performance weakens with age with domain-general competences becoming more relevant. They reported supporting data by comparing the performance of 7- to 9-years-old children and adults, showing that the relationship between ANS and a standardized calculation measure only holds in children. Gimbert et al. (2019) tested the specific contribution of the ANS and WM capacity to math achievement before and after the beginning of formal schooling. They found that ANS accuracy was a predictor of math only in 5-year-old children (r = .34) whereas WM capacity better explained math competence in 7-year-olds pupils (r = .38).

Similarly, many previous findings suggest that mathematical skill is strongly related to WM measures (e.g., Caviola et al., 2014; Geary, 2011; Szűcs et al., 2014). Several models of WM capacity have been proposed, with these varying according to the type of information being manipulated (verbal or spatial; Baddeley, 1986, 2000), or the degree of required cognitive control, ranging from low (i.e., short-term memory, STM) to high level of cognitive control (i.e., WM; Engle, 2010; Cowan 2014). Many recent meta-analyses investigated relationships between WM

measures and mathematics achievement tests in both typically and atypically developing populations (Friso-Van Den Bos, Van Der Ven, Kroesbergen, & Van Luit, 2013; Peng et al., 2016; Peng, Wang, & Namkung, 2018; Szűcs, 2016). For example, Peng et al., (2016) reported a correlation of r = .38 between mathematical achievement scores and composite WM scores and somewhat lower correlation between separate measures of verbal and spatial WM and math achievement (r = .30 and r = .31 respectively; Peng et al., 2016; for similar results see also Friso-Van den Bos, et al., 2013). The verbal WM component seems more involved in the earliest stages of learning, such as counting (Logie & Baddeley, 1987), and the verbal mapping of quantity representations (Menon, 2016; Raghubar, Barnes, & Hecht, 2010). Spatial STM and WM seem to provide a mental workspace for manipulations and are often found to be weak in children with mathematical learning disabilities (Ashkenazi, Rosenberg-Lee, Metcalfe, Swigart, & Menon, 2013; Mammarella, Caviola, Cornoldi, & Lucangeli, 2013; Mammarella, Caviola, Giofrè, & Szűcs, 2018; Passolunghi & Mammarella, 2010; Szűcs et al., 2014; Szűcs, Devine, Soltesz, Nobes, & Gabriel, 2013a).

Szűcs et al. (2014) considered many potential developmental predictors of standardized mathematical performance besides the ANS in 98 ten-year-olds. Using robust bootstrap statistics they found zero-order correlations of 0.26 and 0.25 between math achievement and some ANS and SNC measures, respectively. These results are well within the confidence intervals suggested by Schneider et al. (2017). More interestingly, Szűcs et al. (2014) found that when comparison measures were entered into regression models with verbal and spatial WM measures they were no longer relevant predictors of math achievement (with SNC being a more reliable predictor than ANS). Therefore, their connection to math achievement may be weaker than the connection between math achievement and WM measures (Friso van der Bos, et al. 2013; Peng, et al., 2016). In addition, it seems that math achievement correlates with so many cognitive variables that it is not very surprising or unexpected to find a correlation between math and a randomly picked cognitive construct (Szűcs et al. 2014). These findings point to the importance of studying the context of multiple variables rather than just focusing on isolated relationships between 2-3 constructs.

The present study

We argue that the (developmental) number cognition field lacks and needs high powered developmental studies that consider multiple variables rather than just zero-order correlations of a few variables. Here we report such a study of several derived variables of 7 cognitive constructs from the data of 1254 children of three different age groups. In order to determine whether findings were specific to math achievement we also used reading as an outcome measure. To our knowledge, only two other large sample numerical developmental studies have measured a similarly large number of variables also including measures of the ANS (Lyons, Price, Vaessen, Blomert, & Ansari, 2014; Wei et al., 2012; see Discussion).

We considered various potential correlates and models of math achievement and reading decoding in three age groups covering the primary school years (Grades 2, 4, and 6). Due to high statistical power we were able to estimate *effects*, their *relative* importance and *specificity* to math *precisely* with high *time resolution* across development. We examined the relationship between math achievement or reading decoding, as a control variable, with both domain specific magnitude comparison measures (ANS and SNC) and domain general measures (verbal and spatial STM and WM). We also controlled for fluid intelligence that has been shown to be a strong predictor of general academic performance, including math achievement (e.g., Alloway & Passolunghi, 2011; Colom, Escorial, Shih, & Privado, 2007; Giofrè Mammarella, & Cornoldi, 2014). Further, by regression models we tested the relative contributions of magnitude comparison, STM and WM measures as predictors of math achievement, separately for each grade. Following Geary (2013) and Inglis and colleagues' (2011) theoretical models, it could be expected that during development ANS variables (potentially linked to an evolutionarily based 'primitive' ability) will become weaker correlates of math achievement whereas WM processes (linked to mental manipulations) will become stronger correlates of math performance. To examine whether regression models were specific predictors of math achievement the same regression models were fitted to reading decoding. Importantly, to disentangle methodological issues related to measures of ANS previously used in research studies, we considered often neglected visual display parameter confounds (the congruity of numerical and visual display information) in ANS tasks and computed fits for various model combinations.

Methods

A full Methods section is available in the Supplementary Methods, here we present an abbreviated version.

Participants

Table 2 shows count, grade, age and gender data for the 1254 children who were included in analyses. The analyzed sample size may be smaller for specific analyses because of the constraints linked to the metrics calculated on ANS task. Therefore, for each analysis the sample size is clearly reported. Data was collected in schools located in northeastern Italy. The study received ethical permission from the Psychology Research Ethics Committee of the University of Padova. Written, informed consent of parents or guardians was obtained before testing.

Materials

Academic achievement and intelligence measures

Math achievement was assessed using the standardized AC-MT batteries (Cornoldi & Cazzola, 2004; Cornoldi, Lucangeli, & Bellina, 2012). Both the batteries are comprised of different subtests targeted at different maths learning components. In particular, the following subtests were selected: judging magnitude task (i.e., choosing the larger of set numbers); approximate calculation (i.e., detecting the approximate result of a problem series); retrieving combinations and numerical facts; forward or backward counting knowledge; complex mental and written calculation; transcoding (writing in Arabic format a series of numbers spoken aloud by the experimenter). All the single subtest accuracy scores were summed to create a standardized composite score of maths achievement.

Reading achievement was assessed with standardized tasks derived from the battery for the assessment of Developmental Dyslexia and Dysorthographia-2 (DDE-2, Sartori, Job, & Tressoldi, 2007). Children completed two subtests requiring them to read a few lists of real and pseudo-words. These tasks provide a total both for reading speed and reading accuracy.

The Cattell Culture Fair Intelligence Test (Cattell & Cattell, 1981) was administered to measure nonverbal reasoning (fluid intelligence). The score is the sum of correct answers across all the subtests.

Magnitude representation/comparison tasks

A non-symbolic magnitude comparison task measured ANS. Children compared the numerosity of two sets of black dots on a white background and indicate which set contains more dots by pressing the button on the side of the larger set (Szűcs, et al., 2013b). Ten different number pairs were used with 5 different ratios and their reciprocals (ratios 0.5, 0.62, 0.74, 0.81

and 0.88). In half of trials numerical and visual information (specifically, convex hull) was congruent, in the other half of trials this information was incongruent. As in Szűcs et al., 2013b, we computed overall task accuracy and solution times as well as the so-called *w*, for the overall trials and for both congruent and incongruent trials only.

The SNC task, previously used by Szűcs, et al. (2014), measures the ability of people to compare the relative magnitude of digits. During the task, participants were presented with single Arabic digits and had to decide whether the presented digits were smaller than 5 (indicated by pressing a button with their left hand) or larger than 5 (indicated with a right-hand button press). In line with the recent literature, we calculated accuracy, RTs, and distance effect measures.

Working memory task

Two simple memory span tasks assessed STM. The word span task required the sequential verbal repetition of a series of words, proceeding from the shortest series to the longest. A matrix span task measured spatial STM, where children were asked to memorize and recall the positions of blue cells that appear briefly in different positions on a visible grid in the centre of the screen. WM was measured by a verbal and a spatial dual task (Giofrè, Mammarella, & Cornoldi, 2013) that required participants to concurrently perform a primary and secondary task requiring them to manipulate and recall stimuli. The verbal WM material consisted of a number of word lists. The word lists were organised into sets of different length (i.e., from 2 to 6 words to recall). The primary task required recall of the last word in each list, in the right order of presentation, while the secondary task was to press the space bar when children heard an animal noun. The spatial WM task was comprised of sets of white/grey matrixes in which a black dot would appear and disappear on the grid. Dot sequences were organised into sets of different length (i.e., from 2 to 6 dots to recall). The primary task was to recall the last position of the dot (i.e., the third position for each set). In the secondary task children had to press the spacebar if the dot was presented on a grey cell. The partial credit score was computed for all the four tasks (Conway, et al., 2005; Giofrè & Mammarella, 2014).

Procedure

Each child was tested in their school over 3 sessions between the end of January and May 2018. Children were tested once in groups and twice in individual sessions. Group sessions were used for administering the Fluid intelligence task and some subtests from the Maths achievement batteries (according the administration manual). Children completed the tests under test-like conditions: the children's tables were separated and they were discouraged from speaking with

neighbours. The order of test administration was counterbalanced across classes. Following the group session, two individual sessions, lasting approximately 50 minutes each, were used for administering the Reading tasks, the remaining tasks of the Math batteries, and all the computerized tasks (two Magnitude comparison tasks and four Working memory tasks). Both paper-and-pencil and computerized tasks were equally divided and counterbalanced across the two sessions.

Statistics

At the outset, a series of zero-order and partial correlations, controlling for the fluid intelligence task (hereafter: Cattell), were computed separately for each grade. In order to quantify the evidence for the absence of a correlation, we also report Bayes factor values for the correlation coefficients following the procedure from Wagenmakers, Verhagen, & Ly (2016).

In order to determine the importance of individual predictors, simultaneous linear regression was used throughout this study. We fitted four regression models (Models 1, 2, 3, and 4) to the composite score of math achievement. To examine whether predictors were specific to math we fitted the same four regression models to the reading outcome measure. As we employed standardized scores for the math and reading outcome measures, and non-standardized scores for the predictor variables, we analyzed each grade separately.

There were predictors in common across the four models as well as predictors unique to each model. The predictors that were common to all the models were the four metrics derived from the SNC task (i.e., SNC accuracy, SNC RT and [SNC] distance effect accuracy and RT), the two STM (verbal and spatial), and two the WM scores (verbal and spatial). The differences between the four models (that is, the predictors that were unique to each of the four models) were in the ANS task variables that were included in each: **Model 1** included the Weber fraction computed across all trials. **Model 2** contained two Weber fraction variables, computed separately for congruent and incongruent trials. **Model 3** included ANS accuracy computed across all trials. Finally, **Model 4** included ANS accuracy computed separately for congruent and incongruent trials. **Model 1** and **Model 2** include fewer cases because it is not possible to compute the Weber fraction (the model does not converge and produces arbitrarily large *w* values) for participants that had accuracy scores below 55%.

We adopted a model comparison approach that allowed us to compare our maximal model containing all the predictors (**Version A**) with three theoretically motivated trimmed versions that only contained a subset of the predictors (**Version B** and **Version C** and **Version D**). **Version A**

of each model contained all the predictors. **Version B** of each model dropped the predictors thought to tap into symbolic and non-symbolic magnitude representations (ratio and distance effects) but kept overall SNC accuracy and overall SNC RT as predictors. **Version C** of each model dropped all the predictors derived from the SNC and ANS tasks—that is, **Version C** of the model contained only the predictors derived from the STM and WM task. Finally, **Version D**, dropped all the STM and WM measures from the maximal model. That is, it only contained measures derived from the SNC and ANS tasks. Note that if we include a more extensive set of models including a version that contains only the ANS task measures and a version that contains only the ANS task measures and the symbolic distance effect measures then these models are never selected. A schematic description of the four versions (Version A–D) is shown in **Table 1.**

To perform **model selection** between competing versions and to pick the preferred version we computed two metrics (i.e., AIC [Akaike, 1974] and Cross-validation (CV) mean square error [implanted in DAAG package in R [Maindonald and Braun, 2015]). In summary, we reported the regression fits for the model that both does a good job of explaining the variance in the outcome variable while also containing the fewest number of predictors possible. Therefore, we report a total of *12 regression models* (1 preferred version times 4 Models times three grades) – the preferred version (whether that be version A, B, C, D) of Model 1, 2, 3, and 4 for each grade. Our primary interest was in the significant predictors for the most parsimonious version for each of the four models in each of the grades.

Results

Table 2 reports demographic information, achievement tests, and intelligence scores. **Table 3** shows magnitude comparison and WM results. **Supplementary Table S1 and S2** show the pair-wise mean difference and 95% confidence intervals for between-grade differences.

Zero-order correlations

Zero-order correlations (by grade) between the math and reading composite scores and the other measures are shown in **Supplementary Figures S1 and S2,** respectively. Because our sample size is very large, even very weak correlations (i.e., $r < .11$) can be statistically significant and, therefore, we report Bayes factor (*BF*) values as a measure of evidence in favor of a correlation (BF_{10}) or the absence of a correlation (BF_{01}). Supplementary Tables S3 (and S5 for partial correlations) and S4 (and S6 for partial correlations) provide details of *r* values, 95% CI, and *BF* values, with math and reading composite scores respectively. Additionally, heatmaps indicating the level of evidence in favor of a correlation or in favor of the null are shown in Figure S3. Table S7 reports zero order correlations (by grade) between Cattell (IQ) and all the other measures.

In order to help the overview of the large number of results we present correlation results and *BF*s for math and reading in a simplified form in **Tables 4 and 5**, respectively. In particular, we reported whether null (0) or alternative (1) hypotheses were supported by Bayesian analysis and the magnitude of the Bayes Factors. The larger the absolute value of the number, the stronger is the evidence (0=weak; 1=substantial; 2=strong; 3=very strong; 4=decisive). This gives a composite where, for example, 0-2 would indicate strong evidence for the null and, for example, 1+1 would indicate substantial evidence for the alternative.

For the correlation between the Weber fraction and math scores, we found weak evidence in favor of a correlation for Grade 2 and Grade 4 and decisive evidence in favor of a correlation for Grade 6, when all trials were examined together. When only congruent trials were examined, we found substantial and very strong evidence in favor of a correlation for Grade 2 and 6, but strong evidence for the absence of a correlation in Grade 4. When only incongruent trials were examined, we found substantial evidence for the absence of a correlation in all Grades. A similar pattern of results (only with slight differences in the strength of evidence) was observed for the correlations between ANS accuracy and math scores. This is unsurprising given that ANS accuracy and the Weber fraction are, by definition, highly correlated (see Szűcs et al. 2013b).

For the ANS RT measure, we found substantial to strong evidence for the absence of a correlation in Grade 2 and Grade 4, when all trials were examined, when congruent trials were examined alone, and when incongruent trials were examined alone. For Grade 6, however, weak to substantial evidence was found in favor of a correlation when all trials were combined or

congruent trials were examined alone. When incongruent trials were examined alone, we found weak evidence for the absence of a correlation.

In contrast to the results for ANS RT and accuracy, which were mixed, we found strong to decisive evidence in favor of a correlation between SNC RT and accuracy and math score for all Grades. However, the picture was more complex for the (SNC) distance effect measures (both RT and accuracy). For the correlation between the distance effect (accuracy) and math scores, we found weak evidence for no correlation in Grade 2, weak evidence for a correlation in Grade 4, and strong evidence for no correlation in Grade 6. For the (SNC) distance effect RT measure, we found strong evidence for no correlation in Grade 2, and weak to substantial evidence for a correlation in Grade 4 and 6.

For the STM and WM tasks, we found decisive evidence in favor of a correlation between all measures and math scores in all grades except in one case. This was the correlation between spatial STM and math scores in Grade 4, which only provided weak evidence in favor of a correlation.

For the correlations between the reading scores and the ANS measures (Weber fraction and accuracy), we generally found weak to strong evidence for the absence of a correlation, except in a few cases. These were the weber fraction (all trials and congruent trials) in Grade 6, ANS accuracy (congruent trials) in Grade 6, and ANS RT (all trials, congruent and incongruent trials) in Grade 2, where we found weak to strong evidence in favor of a correlation with reading rate.

For correlations between reading scores and the SNC task, we found weak to decisive evidence in favor of a correlation for SNC RT. For SNC accuracy, we found weak to substantial evidence in favor of no correlation for Grade 4 and Grade 6, while for Grade 2 we found weak evidence for a correlation. For SNC distance effect measures (both RT and accuracy), we found weak to strong evidence for no correlation.

Finally, we found substantial to decisive evidence for a correlation between verbal STM and reading scores and verbal WM and reading scores across all Grades. For the spatial STM measure we found strong evidence in favor of correlation in Grade 2; conversely in Grade 4 and 6 we found weak to substantial evidence for the absence of a correlation. For the spatial WM measure we found decisive to strong evidence in favor of a correlation for Grade 2 and Grade 6 and weak evidence in favor of a correlation in Grade 4.

Regression models

Standardized β values, and 95% confidence intervals, for each of the regression models (full version) are shown in **Figure 1** (math achievement) and **Figure 2** (reading performance) respectively. The standardized β values, and 95% confidence intervals, for the four versions (Version A, B, C and D) of each model for both the math score (math models) and reading score (reading models) are provided in the Supplementary results.

A summary of the regression analysis is shown in **Table 6**, while the complete regression tables including standardized β values, and 95% confidence intervals, for the three versions of each models are provided in the supplementary results (Table S8–Table S15)

Model 1: Weber fraction computed for all trials.

For the mathematics model, cross-validation selected Version B for all three Grades. Once all the data was fitted to the preferred version, SNC accuracy, spatial WM, verbal WM, and verbal STM were significant predictors in *all grades*. In addition, SNC RT was a significant predictor in Grade 6.

The same cross-validation procedure was used to select the preferred specification of the reading model. Version B was selected in Grade 4 and Grade 6, and version C was selected in Grade 2. Once all the data was fitted to the preferred specification, verbal WM was a significant predictor in *all Grades*. In addition, verbal STM and SNC RT were significant predictors in Grade 4 and Grade 6 while spatial WM was a significant predictor in Grade 2.

Importantly, verbal WM was a significant predictor of *both* reading and math across all three grades, suggesting that it is tracking a cognitive capacity that is not specific to mathematics. Similarly, verbal STM was a significant predictor for both reading and math in Grade 4 and Grade 6, while spatial WM, and SNC RT were significant for both reading and math in Grade 2 and Grade 6, respectively. In contrast, SNC accuracy was significant only for mathematics in *all three grades* suggesting that it is specific to math. Similarly, spatial WM also appeared to be specific to math, at least in Grade 4 and Grade 6.

Model 2: Weber fraction computed separately for congruent and incongruent trials

For the mathematics model, cross-validation selected Version A for Grade 2, Version C for Grade 4, and Version B for Grade 6. Once the preferred model was fit to the entire dataset, spatial WM and verbal STM were significant in *all Grades*. In addition, SNC accuracy, SNC RT, and verbal WM were significant predictors in Grade 2 and Grade 6, and the weber fraction (congruent trials only) was a significant predictor in Grade 2.

Version B of the reading model was selected in all three Grades. Once the preferred specification was fit to the entire dataset SNC RT, verbal WM, and verbal STM were significant predictors in Grade 4 and Grade 6, while spatial WM was a significant predictor in Grade 2.

Several predictors appeared to be specific to math in at least one grade. These were the weber fraction for congruent trials (Grade 2), SNC accuracy (Grade 2 and Grade 6), verbal STM (Grade 2), verbal WM (Grade 2), and spatial WM (Grade 4 and Grade 6).

Model 3: ANS accuracy computed for all trials

Cross-validation selected version A of the mathematics model in Grade 6, with version B selected in Grade 2 and Grade 4. Once all the data was fit to the preferred specification of the model, SNC accuracy, spatial WM, verbal WM, and verbal STM were significant predictors of math in *all three Grades*. In addition, SNC RT was a significant predictor of math in Grade 4 and Grade 6, and the SNC distance effect (accuracy) was a significant predictor in Grade 6 only.

Cross-validation selected version C of the reading model in Grade 2 and version B in Grade 4 and Grade 6. Once all the data was fit to the preferred version in each Grade, verbal WM was a significant predictor of reading in all three Grades. In addition, SNC RT and verbal STM were significant predictors of reading in Grade 4 and Grade 6 while spatial WM was a significant predictor of reading in Grade 2.

Comparing the mathematics models and the reading models we can see that verbal WM was a significant predictor of both reading and math across *all three Grades*. In addition, SNC RT was a significant predictor of reading and math in Grade 4 and Grade 6, suggesting that it is tracking a cognitive process that is not specific to math. Similarly, verbal STM was significant for both reading and math in Grade 4 and Grade 6 again suggesting that these variables track cognitive capacities not specific to math. Of the predictors that were specific to math, SNC accuracy was significant across *all three Grades* and spatial WM was significant in Grade 4 and Grade 6 (while being shared between reading and math in Grade 2).

Model 4: ANS accuracy computed separately for congruent and incongruent trials.

Cross-validation selected version B of the mathematics model for Grade 2, while version A was preferred in Grade 4 and Grade 6. Once all the data was fit to the preferred version in each Grade spatial WM and verbal STM were significant predictors of math in *all three grades*. Verbal WM and SNC accuracy were also significant predictors in Grade 2 and Grade 6. Finally, predictors that were significant in only one grade included the SNC distance effect (accuracy) in Grade 6, ANS accuracy (incongruent trials only) in Grade 4, and SNC RT in Grade 6.

For the reading outcome variable, cross-validation selected version C for Grade 2, while version B was preferred in Grade 4 and Grade 6. Once all the data was fit to the preferred version in each Grade, verbal WM was a significant predictor of reading across *all three Grades*, verbal STM and SNC RT were significant predictors in Grade 4 and Grade 6, and spatial WM was a significant predictor in Grade 2.

Of the significant predictors, verbal WM (Grade 2 and Grade 6) and verbal STM (Grade 4 and Grade 6) were significant predictors of both reading and math in at least two Grades. In addition, SNC RT was a predictor of both reading and math in Grade 6 and spatial WM was a predictor of both reading and math in Grade 2. Of the predictors that were specific to math, spatial WM was significant for two Grades (Grade 4 and Grade 6), as well as SNC accuracy (Grade 2 and Grade 6), while the SNC distance effect (accuracy; Grade 6), ANS accuracy (incongruent trials; Grade 4), and verbal STM (Grade 2) where significant in only one Grade.

Regression summary

Looking across all four model specifications a few general patterns can be observed. The predictors that were specific to math were symbolic accuracy and spatial WM (although in Grade 2, spatial WM was also a predictor of reading). Non-specific predictors included verbal STM and verbal WM. Similarly, symbolic RT was often found as a significant predictor for both reading and mathematics, again suggesting a lack of specificity. Finally, the measures derived from the ANS task (non-symbolic accuracy, RT, or Weber fraction) were found to be significant predictors of math in only one occasion. This was for the non-symbolic accuracy (incongruent trials only) in Grade 4.

Discussion

Our cross-sectional study aimed at clarifying the relative importance of the ANS, SNC and verbal and spatial STM and WM for mathematics achievement in 1254 Grade 2, 4 and 6 children. The large cohort of participants assured high power and low false report probability and allowed us to determine effect sizes precisely. We included various measures of 7 important constructs underlying mathematics performance. Hence, we could test relationships in the context of potentially important alternative variables rather than in isolation. We computed zero-order and partial correlations controlling for fluid intelligence and we determined the relative weight of each variable in regression models. Reading decoding served as control outcome measure to test

whether findings were specific to mathematics. Bayesian analysis allowed us to provide probabilities for null and alternative hypotheses.

Zero-order correlations between math and *w* and between math and accuracy computed for all trials and/or congruent trials was generally weak, expect in Grade 6 where support for a correlation was stronger ($0.18 \leq |r| \leq 0.01$). Zero-order correlations are in-line with those obtained by Halberda et al., (2012) who found a zero-order correlation of $r = -0.16$ between *w* (computed without considering visual confounds) and the self-reported Scholastic Aptitude Test (SAT) scores in 458 adults. Halberda et al. (2012) also reported correlations of ($-0.23 \leq r \leq -0.13$) between *w* and self-reported math expertise in mostly adult age groups (see Table 1 in Halberda et al., 2012). Most relevantly, Halberda et al. (2012) reported $r = -0.13$ between *w* and self-reported math expertise in 994 children aged 11-17 years. We conclude that our results replicate the previously reported zero-order correlations from the large study of Halberda et al. (2012).

Two other large studies measured larger effect sizes than us and Halberda et al. (2012). Lyons et al., (2014; N=1391; 201-253 children in each of 6 grades) used only trials where numerical and visual information was incongruent and used a mental arithmetic test of 50 additions and 50 subtractions as dependent measure. They reported correlations of $0.143 \leq r \leq 0.321$ between ANS accuracy (32 trials in one-digit and 32 trials in two-digit range) and mental arithmetic. However, based on regression results they ultimately concluded that ANS performance was not a significant predictor of mental arithmetic in any of the grades. Another large study (Wei et al., 2012; N=1556) reported $r = -0.39$ and $r = -0.3$ between a ANS task (36 trials per participant) and subtraction and multiplication performance, respectively. However, in this study dot numbers ranged between 5-12 and no display timeout is mentioned. Hence, counting may have been used for responding (note that there is no theoretical reason for a multiplication task to be related to ANS task; Dehaene, 1997).

The most apparent reason for the discrepancy between our results and Halberda et al. (2012) vs. Lyons et al. (2014) and Wei et al. (2012) may be the different nature of math outcome tests. Both us and Halberda et al. (2012) relied on math curriculum tests and the wider concept of 'self-declared math expertise' (Halberda et al., 2012) whereas the other two studies used narrow mental arithmetic tests. Hence, it seems that curriculum tests and wider math competence have lower correlations with ANS measures than mental arithmetic tests. This conclusion is also supported by meta-analyses. First, overall these analyses found larger effect sizes than us and Halberda et al. (2012) (For ANS task measures overall: $0.22 \leq r \leq 0.24$; Chen and Li, 2014; Fazio et al. 2014;

Schneider et al., 2017. For child groups with similar ages as tested here r = 0.280. Correlation for w = 0.315; Schneider et al. 2017). Second, Schneider et al. (2017) also separated studies by their outcome measures and reported that mental arithmetic tests had substantially larger correlation with w than curriculum-based measures (0.378 vs. 0.205). In fact, the meta-analytic estimate for curriculum measures is close to the range and maximum values found both by us and Halberda et al. (2012). Considering that the data from Schneider et al. also includes adult studies and meta-analyses are subject to effect size exaggeration (Ioannidis, 2008; Szűcs & Ioannidis, 2017b) the larger meta-analytic estimate is not surprising.

The above observations and the good agreement between our study (160 trials per participant in a one-on-one test) and Halberda et al. (2012; 300 trials per participant in an online test) and Schneider et al. (2017) suggest that in ANS tasks the typical w vs. math curriculum test zero-order correlation effect size in school-aged children is in the range detected here. Notably, this effect size is very small, equivalent to an r^2 value of at most $(-0.18)^2 = 0.0324$. That is, less than 4% of the variability in children's math scores is predicted by their *ANS task performance*. This small effect size renders w unsuitable for individual diagnosis of, for example, children with developmental dyscalculia (Szűcs, et al., 2013a).

We have separately analyzed congruent and incongruent trials of the ANS task. There was no reliable relationship between math and w computed from incongruent trials. So, visual cues can sufficiently disturb performance in the ANS task so that it loses construct validity in terms of claims regarding a general relationship with numerical skills. The above also suggests that w vs. math correlations arose fully from the influence of congruent trials. Importantly, congruent trials do not provide a 'pure' measure of the ANS either as in these trials visual cues are positively correlated with numerical information. Hence, both better visual cue discrimination and better numerical decisions can explain correlations in these trials. In contrast to our study, Lyons et al. (2014; see above), Fuhs and McNeil (2013; N=103 pre-schoolers of 3.7-5.9 years of age; r = 0.23), and Gilmore et al., (2013; N=80 children aged 4.7-11.9 years; r = 0.55) reported significant correlations between math and ANS task performance in incongruent trials only.

ANS task performance was related to reading performance in Grade 2 ($r \sim -0.18$). However, when fluid intelligence was taken into account no measures of w showed reliable relations with math in any grades ($-0.11 \leq r \leq +0.01$). In addition, when w vs. math correlations were present we also found correlations of similar magnitude not only between w and fluid intelligence but between w and spatial STM and WM as well. Similarly, Szűcs et al., (2014) found that the best

correlates of *w* were sustained attention, phonological decoding and a STM task. We conclude that the *w* vs. math relation is heavily influenced by some components of general cognitive processes, such as fluid intelligence, executive functions and cognitive control abilities (Gilmore et al., 2013; Leibovich et al., 2017; Xenidou-Dervou et al., 2018). Hence, at least in the school age populations tested here, the *w* vs. math relation is likely spurious and whereas ANS tasks and math may correlate with similar cognitive factors the two likely do not have any causal connection. Indeed, a mediation analysis (Supplementary Figure S4) suggests that the relationship between several of our measures and math is at least partially mediated by fluid intelligence.

Our results were similar for *w* and ANS task accuracy. This is not surprising because *w* is a direct non-linear function of accuracy data (see Szűcs, et al., 2013b for details). ANS RT never showed correlation with math.

SNC accuracy and SNC RT were reliably correlated with math in all grades (accuracy: $0.16 \leq r \leq 0.29$; RTs: $0.17 \leq r \leq 0.37$) even when fluid intelligence was considered, except the partial correlation for RT in Grade 4. The distance effect measures were not correlated with math in Grade 2 (RT and accuracy) or Grade 6 (accuracy) and they showed *weak* (according to Bayes Factor values) correlations with math in other grades. All but one of the correlations with the distance effect was eliminated once fluid intelligence was considered. We conclude that in school-aged children it is unlikely that SNC shows a relation to math because it has a link to the ANS. Assuming an ANS link is neither supported by the lack of strong correlations between ANS measures and math, nor by the lack of strong relations between the SNC distance effect and math. The most likely possibility is that the weak distance effect vs. math correlations are due to some shared variance with decisional abilities (Van Opstal, Gevers, De Moor, & Verguts, 2008; Olivola & Chater, 2017) that also rely on general fluid intelligence (similar to our above conclusion about the ANS task). Our data also shows that unlike the distance effect, overall SNC accuracy and RT are weak but reliable correlates of math. This is line with the conclusions of other investigators (De Smedt et al., 2013; Lyons et al., 2014; Xenidou-Dervou, Molenaar, Ansari, van der Schoot, & van Lieshout, 2017) and with the meta-analysis of Schneider et al. (2017) who also reported that SNC was a stronger correlate of math than ANS measures.

SNC RT was correlated with reading decoding rate in all grades ($-0.14 \leq r \leq -0.21$) even when considering fluid intelligence, except in Grade 2. This finding is in line with the data of 98 children from Szűcs et al. (2014). Vanbinst et al., (2016) also reported SNC vs. reading correlations of similar effect size in a cross-sectional-longitudinal study of 74 third grade children

(-0.18≤r≤-0.22). However, their lower powered study could not identify the correlations as statistically significant, so they argued that SNC was a domain-specific predictor of arithmetic. Notably, this argument assumed that 'no statistical significance' implies the lack of a relationship (accepting the null hypothesis) which is an invalid conclusion. In contrast, our Bayesian analysis suggested that even intelligence controlled correlations with reading were 'strong' to 'very strong' in Grades 4 and 6. Hence, at least some aspects of SNC do not seem number specific, perhaps due to the involvement of general symbol processing ability, for example symbol–referent processing (Grabner, Ansari, Koschutnig, Reishofer, & Ebner, 2013; Grabner, Reishofer, Koschutnig, & Ebner, 2011; Szűcs et al., 2014).

In contrast to the scattered nature of evidence provided by the ANS task, results were clear cut for measures of verbal (0.27≤r≤ 0.34) and spatial (0.26≤r≤ 0.41) memory. Most zero order and partial correlations showed strong to decisive evidence for a link between memory measures and math achievement. The only exception was spatial STM, which showed a weak correlation in Grade 4 (r= 0.13) and a decisive link in Grade 6 (r= 0.23), but these correlations were not reliable when intelligence was considered. In line with our findings, recent meta-analyses concluded that all WM components are equally strongly associated with math performance (Friso-Van Den Bos et al., 2013; Peng et al., 2016; Szűcs, 2016). Some differences between results can be attributed to the variability of WM tasks in studies as well as to developmental changes in general cognitive resources (Meyer, Salimpoor, Wu, Geary, & Menon, 2010). Similarly to us, others have reported that the link between spatial STM and math varies with age/school grades and is less robust than links between math and WM tasks (Holmes & Adams, 2006; Rasmussen & Bisanz, 2005). In addition, Li and Geary, (2013) also found that that spatial STM had a stronger link to math achievement in older children. Notably, the current literature does not yet allow for the clear characterization of the developmental progression of the links between various WM tasks and math (see e.g. Szűcs, 2016).

Verbal STM and WM were associated with reading in all Grades even after controlling for intelligence. However, spatial WM was not associated with reading in Grades 4 and 6 when controlling for intelligence. This observation is in-line with regression results and is further discussed below.

When we analyzed the data with regression models that tested w (Model 1, no congruency factor; similar to those used by other studies (e.g. Halberda et al., 2012; Libertus, Feigenson, & Halberda, 2011, 2013; Sasanguie et al., 2012, 2015), ANS task variables were never significant

predictors of math achievement. SNC accuracy (all 3 grades), SNC RT (grades 2 and 4), spatial WM (grades 4 and 6) were specific predictors of math. SNC RT was a shared predictor with reading decoding in Grade 6. Spatial WM was a shared predictor with reading in Grade 2 and verbal STM and verbal WM were shared predictors with reading decoding in all grades. Practically the same results were obtained by Model 3, that used ANS task accuracy rather than *w* as a predictor (and included more children). Again, because *w* is derived from accuracy data the similarity in findings can be expected.

Models 2 and 4 separated the congruent and incongruent trials of the ANS task. ANS task measures became specific predictors only twice across all models. In one case this happened using Model 2 when computing *w* from *congruent trials* in Grade 2. In the other case, this happened using Model 4 where ANS task RT from *incongruent trials* was a predictor in Grade 4. Model 2 found that SNC accuracy (Grades 2 and 6) and spatial WM were specific predictors of math achievement in two grades (Grades 4 and 6). Model 4 confirmed the spatial WM findings and showed SNC to be a specific predictor of math in Grade 6. Both Models 2 and 4 suggested that verbal STM and verbal WM were shared predictors of math and reading.

Overall, the best specific predictors of math achievement were SNC accuracy (shown in variable grades) and spatial WM (consistently shown in Grades 4 and 6). Our ANS task related findings are in line with the reviews of De Smedt et al., (2013) and with Schneider et al., (2017). It is noteworthy that SNC is more similar to mathematical competence measures than most other measures (it includes symbolic digits and the smaller/larger numerical operations). Hence, 'transfer' pathways are much shorter between this task and math than between other cognitive measures. In other words, there is probably much larger a priori overlap in the cognitive processes behind SNC tasks and math outcome measures than in the case of other cognitive variables.

We found that the most specific domain-general predictor of math was spatial WM. This finding is in agreement with previous studies with typically developing children (Bull, Espy, & Wiebe, 2008; Caviola et al., 2014) and with children with developmental dyscalculia (Mammarella et al., 2018; Passolunghi & Mammarella, 2010, 2012; Szűcs, 2016; Szűcs, et al., 2013a). A likely possibility is that spatial WM provides an important mental workspace for maintaining and evaluating spatial relations that play a role in mathematics but not in reading (Giofrè, Donolato, & Mammarella, 2018; Szűcs et al., 2014). Results suggest that the importance of these spatial relations increases from earlier to later school grades.

It is not surprising that verbal STM and verbal WM were shared predictors of both math achievement and reading decoding (Berg, 2008; Compton, Fuchs, Fuchs, Lambert, & Hamlett, 2012; Swanson, 2017). In fact, verbal WM has been consistently found to be related to general academic outcomes (Berg, 2008; Bull et al., 2008; Gathercole, Pickering, Knight, & Stegmann, 2004). Previous studies noted the role of verbal memory in encoding and retaining verbal numerical information used for specific math tasks such as counting and/or retaining interim solutions during complex mental calculation (e.g., Bull et al., 2008; Gathercole & Pickering, 2000; Gathercole et al., 2004; Swanson & Sachse-Lee, 2001). Verbal WM can support verbal task-solution strategies (i.e., subvocal rehearsal/retention) and direct retrieval of arithmetic facts from long-term memory (Ashcraft, 1982; Holmes & Adams, 2006). These results are also supported by research on mathematical difficulties: the high comorbidity between math and reading difficulties is well-known and may be explained by co-occuring modality-specific verbal/phonological impairment (Szűcs, 2016).

In our very large sample we could not assess other domain-general factors such as some executive function or cognitive control measures that are related to math achievement (Bull, et al. 2008) and contribute to performance in ANS tasks (Leibovich, et al., 2017; Szűcs et al., 2014) and in WM tasks (Kane & Engle, 2002; McCabe et al., 2010). For this reason, future studies should extend the range of domain-general skills considered. It would also be advantageous to have different curriculum based and standardized measures of math and reading achievement as different outcome measures may have different correlations with cognitive variables as we discussed above. Similarly, including additional domain-specific tasks, for example assessing the mapping between ANS and SNC, may help us to understand developmental change (e.g., Gimbert., et al., 2019). Additionally, it remains to be seen whether ANS task performance is more related to math in pre-school age groups (vanMarle, Chu, Li & Geary, 2014). Our findings were derived from the assessment of school-aged children and cannot be generalized to early developmental periods. However, there have also been several negative results about the importance of ANS for preschool periods (Fuhs & McNeil, 2013; Kolkman, Kroesbergen, & Leseman, 2013; Lyons, Bugden, Zheng, De Jesus & Ansari, 2018; Sasanguie, Göbel, Moll, Smets, & Reynvoet, 2013; Szũcs, Soltész, Jármi, & Csépe, 2007). Thus, the importance of ANS for preschool populations cannot be taken for granted. We suggest that further research efforts should be targeted at whether the ANS does play a (causal) role in early number development, by collecting large samples assuring high power and low false report probability.

Conclusions

Replicating the outcomes from a similar large study (Halberda et al. 2012) we found weak *zero-order* correlations between some ANS measures and math achievement. However, we also found that correlations relied on trials where numerical and visual information were positively correlated and effects ceased to be reliable once fluid intelligence was considered. Similar to previous findings ANS measures correlated with various cognitive variables and they never became significant predictors of math when other variables were included in regression models (see Szűcs et al. 2014; Lyons et al. 2014). Hence, we conclude that, at least in school age populations, ANS measures are spurious correlates of curriculum level math achievement and they are unlikely to reflect any causal connections between ANS and math achievement. The low predictive power of the ANS task makes it unsuitable for diagnosing complex conditions such as developmental dyscalculia and make it unlikely that ANS training could result in curriculum level benefits (see Szűcs & Myers, 2017, for an analysis of ANS training studies).

We found that SNC accuracy was a reliable and largely specific correlate of math achievement. This relation is unlikely to draw on the ANS. Rather, it may reflect human specific math or more general symbol processing ability. We found that verbal WM performance supports both reading and math achievement. In contrast, spatial WM is an increasingly specific correlate of math, the specific relation becoming stronger in older children (Grades 4 and 6 here). Spatial WM likely provides a mental workspace utilized in math but not in reading performance (Szűcs et al., 2014). Nevertheless, considering that to date mere spatial WM training proved ineffective in improving math performance (Melby-Lervåg, Redick, & Hulme, 2016) the exact links and impact mechanism between spatial WM and math performance need to be uncovered.

Data Availability Statement

The data that support the findings of this study are openly available in a GitHub.com repository at http://dx.doi.org/10.17605/OSF.IO/SEP78.

References

Akaike, H. (1974). A New Look at the Statistical Model Identification (pp. 215–222). Springer, New York, NY. doi: 10.1007/978-1-4612-1694-0_16

Alloway, T. P., & Passolunghi, M. C. (2011). *The relationship between working memory, IQ, and mathematical skills in children*. *Learning and Individual Differences* (Vol. 21). doi: 10.1016/j.lindif.2010.09.013

Ansari, D. (2008). Effects of development and enculturation on number representation in the brain. *Nature Reviews Neuroscience*. doi: 10.1038/nrn2334

Ashcraft, M. H. (1982). The development of mental arithmetic: A chronometric approach. *Developmental Review*, *2*(3), 213–236. doi: 10.1016/0273-2297(82)90012-0

Ashkenazi, S., Rosenberg-Lee, M., Metcalfe, A. W. S., Swigart, A. G., & Menon, V. (2013). Visuo–spatial working memory is an important source of domain-general vulnerability in the development of arithmetic cognition. *Neuropsychologia*, *51*(11), 2305–2317. doi: 10.1016/j.neuropsychologia.2013.06.031

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

Barrouillet, P. (2018). Numerical cognition and memory(ies). In A. Henik & W. Fias (Eds.), *Heterogeneity of function in numerical cognition* (pp. 361–386). San Diego: Academic Press.

Berg, D. H. (2008). Working memory and arithmetic calculation in children: The contributory roles of processing speed, short-term memory, and reading. *Journal of Experimental Child Psychology*, *99*(4), 288–308. doi: 10.1016/J.JECP.2007.12.002

Bull, R., Espy, K. A., & Wiebe, S. a. (2008). Short-term memory, working memory, and executive functioning in preschoolers: Longitudinal predictors of mathematical achievement at age 7 years. *Developmental Neuropsychology*, *33*(3), 205–228. doi: 10.1080/87565640801982312

Cattell, R. B., & Cattell, A. K. S. (1981). Measuring intelligence with the culture fair tests. Institute for Personality and Ability Testing [Misurare l'intelligenza con i test "Culture Fair"]. Florence: Giunti OS.

Caviola, S., Mammarella, I. C., Lucangeli, D., & Cornoldi, C. (2014). Working memory and domain-specific precursors predicting success in learning written subtraction problems. *Learning and Individual Differences*, *36*, 92–100. doi: 10.1016/j.lindif.2014.10.010

Chen, Q., & Li, J. (2014). Association between individual differences in non-symbolic number acuity and math performance: A meta-analysis. *Acta Psychologica*, *148*, 163–172. doi:

10.1016/j.actpsy.2014.01.016

Colom, R., Escorial, S., Shih, P. C., & Privado, J. (2007). Fluid intelligence, memory span, and temperament difficulties predict academic performance of young adolescents. *Personality and Individual Differences*, *42*(8), 1503–1514. doi: 10.1016/J.PAID.2006.10.023

Compton, D. L., Fuchs, L. S., Fuchs, D., Lambert, W., & Hamlett, C. (2012). The Cognitive and Academic Profiles of Reading and Mathematics Learning Disabilities. *Journal of Learning Disabilities*, *45*(1), 79–95. doi: 10.1177/0022219410393012

Conway, A., Kane, M., Bunting, M., Hambrick, D. Z., Wilhelm, O., and Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review, 12*(5):769–786. doi: 10.3758/BF03196772

Cornoldi, C., Lucangeli, D., & Bellina, M. (2012). *AC-MT 6-11: Test for Assessing Calculation and Problem Solving Skills* [Test AC-MT 6-11 - Test di Valutazione delle Abilità di Calcolo e Problem Solving]. Trento: Erickson.

Cornoldi C., & Cazzola C. (2004). *AC-MT 11-14: Test for Assessing Calculation and Problem Solving Skills* [Test AC-MT 11-14 - Test di Valutazione delle Abilità di Calcolo e Problem Solving]. Trento: Erickson.

Cowan, N. (2014). Working memory underpins cognitive development, learning, and education. *Educational psychology review, 26*(2), 197-223. doi: 10.1007/s10648-013-9246-y

Cragg, L., & Gilmore, C. (2014). Skills underlying mathematics: The role of executive function in the development of mathematics proficiency. *Trends in Neuroscience and Education*, 3(2), 63–68. doi: 10.1016/j.tine.2013.12.001

De Smedt, B., Noël, M. P., Gilmore, C., & Ansari, D. (2013). How do symbolic and non-symbolic numerical magnitude processing skills relate to individual differences in children's mathematical skills? A review of evidence from brain and behavior. *Trends in Neuroscience and Education*, *2*(2), 48–55. doi: 10.1016/j.tine.2013.06.001

Dehaene, S. (1997). *The number sense*. Oxford: Oxford University Press.

Dietrich, J. F., Huber, S., & Nuerk, H.-C. (2015). Methodological aspects to be considered when measuring the approximate number system (ANS)—A research review. Frontiers in Psychology, 6. doi: 10.3389/fpsyg.2015.00295.

Engle, R. W. (2010). Role of working-memory capacity in cognitive control. *Current Anthropology, 51*, S17–S26. doi: 10.1086/650572

Fabbri, S., Caviola, S., Tang, J., Zorzi, M., & Butterworth, B. (2012). The role of numerosity in

processing nonsymbolic proportions. *The Quarterly Journal of Experimental Psychology*, *65*(12), 2435-2446. doi: 10.1080/17470218.2012.694896

Fazio, L. K., Bailey, D. H., Thompson, C. A., & Siegler, R. S. (2014). Relations of different types of numerical magnitude representations to each other and to mathematics achievement. *Journal of Experimental Child Psychology*, *123*, 53–72. doi: 10.1016/j.jecp.2014.01.013

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*(7), 307–314. doi: 10.1016/j.tics.2004.05.002

Fias, W., & Menon, V. (2013). Multiple components of developmental dyscalculia. *Trends in Neuroscience and Education*, *2*(2), 43–47. doi: 10.1016/j.tine.2013.06.006

Fias, W., Menon, V., & Szűcs, D. (2013). Multiple components of developmental dyscalculia. *Trends in Neuroscience and Education*, *2*(2), 43–47. doi: 10.1016/j.tine.2013.06.006

Friso-Van Den Bos, I., Van Der Ven, S. H. G., Kroesbergen, E. H., & Van Luit, J. E. H. (2013). Working memory and mathematics in primary school children: A meta-analysis. *Educational Research Review*, *10*, 29–44. doi: 10.1016/j.edurev.2013.05.003

Fuhs, M. W., & McNeil, N. M. (2013). ANS acuity and mathematics ability in preschoolers from low-income homes: contributions of inhibitory control. *Developmental Science*, *16*(1), 136–148. doi: 10.1111/desc.12013

Gathercole, S. E., & Pickering, S. J. (2000). Working memory deficits in children with low achievements in the national curriculum at 7 years of age. *British Journal of Educational Psychology*, *70*(2), 177–194. doi: 10.1348/000709900158047

Gathercole, S. E., Pickering, S. J., Knight, C., & Stegmann, Z. (2004). Working memory skills and educational attainment: evidence from national curriculum assessments at 7 and 14 years of age. *Applied Cognitive Psychology*, *18*(1), 1–16. doi: 10.1002/acp.934

Geary, D. C. (2011). Cognitive predictors of achievement growth in mathematics: a 5-year longitudinal study. *Developmental Psychology*, *47*(6), 1539–52. doi: 10.1037/a0025510

Geary, D. C. (2013). Early foundations for mathematics learning and their relations to learning disabilities. *Current Directions in Psychological Science, 22*, 23–27. doi: 10.1177/0963721412469398

Gebuis, T., & Reynvoet, B. (2012a). The interplay between nonsymbolic number and its continuous visual properties. *Journal of Experimental Psychology: General*, *141*(4), 642–648. doi: 10.1037/a0026218

Gebuis, T., & Reynvoet, B. (2012b). The Role of Visual Information in Numerosity Estimation.

PLoS ONE, *7*(5), e37426. doi: 10.1371/journal.pone.0037426

Gilmore, C., Attridge, N., Clayton, S., Cragg, L., Johnson, S., Marlow, N., … Inglis, M. (2013). Individual Differences in Inhibitory Control, Not Non-Verbal Number Acuity, Correlate with Mathematics Achievement. *PLoS ONE*, *8*(6), e67374. doi: 10.1371/journal.pone.0067374

Gimbert, F., Camos, V., Gentaz, E., & Mazens, K. (2019). What predicts mathematics achievement? Developmental change in 5-and 7-year-old children. Journal of experimental child psychology, 178, 104-120. doi: 10.1016/j.jecp.2018.09.013

Giofrè, D., Donolato, E., & Mammarella, I. C. (2018). The differential role of verbal and visuospatial working memory in mathematics and reading. *Trends in Neuroscience and Education*, *12*, 1–6. doi: 10.1016/J.TINE.2018.07.001

Giofrè, D., & Mammarella, I. C. (2014). The relationship between working memory and intelligence in children: Is the scoring procedure important? *Intelligence*, *46*, 300–310. doi: 10.1016/J.INTELL.2014.08.001

Goffin, C., & Ansari, D. (2019). How Are Symbols and Nonsymbolic Numerical Magnitudes Related? Exploring Bidirectional Relationships in Early Numeracy. *Mind, Brain, and Education.* doi: 10.1111/mbe.12206

Grabner, R. H., Ansari, D., Koschutnig, K., Reishofer, G., & Ebner, F. (2013). The function of the left angular gyrus in mental arithmetic: Evidence from the associative confusion effect. *Human Brain Mapping*, *34*(5), 1013–1024. doi: 10.1002/hbm.21489

Grabner, R. H., Reishofer, G., Koschutnig, K., & Ebner, F. (2011). Brain Correlates of Mathematical Competence in Processing Mathematical Representations. *Frontiers in Human Neuroscience*, *5*, 130. doi: 10.3389/fnhum.2011.00130

Halberda, J., Ly, R., Wilmer, J. B., Naiman, D. Q., & Germine, L. (2012). Number sense across the lifespan as revealed by a massive Internet-based sample. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(28), 11116–20. doi: 10.1073/pnas.1200196109

Halberda, J., Mazzocco, M. M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, *455*(7213), 665–668. doi: 10.1038/nature07246

Hohol, M., Cipora, K., Willmes, K., & Nuerk, H.-C. (2017). Bringing Back the Balance: Domain-General Processes Are Also Important in Numerical Cognition. *Frontiers in Psychology*, *8*, 499. doi: 10.3389/fpsyg.2017.00499

Holloway, I. D., & Ansari, D. (2008). Domain-specific and domain-general changes in children's development of number comparison. *Developmental Science*, *11*(5), 644–649. doi: 10.1111/j.1467-7687.2008.00712.x

Holmes, J., & Adams, J. W. (2006). Working Memory and Children's Mathematical Skills: Implications for mathematical development and mathematics curricula. *Educational Psychology*, *26*(3), 339–366. doi: 10.1080/01443410500341056

Inglis, M., Attridge, N., Batchelor, S., & Gilmore, C. (2011). Non-verbal number acuity correlates with symbolic mathematics achievement but only in children. *Psychonomic Bulletin & Review, 18*, 1222–1229. doi: 10.3758/s13423-011-0154-1

Inglis, M., & Gilmore, C. (2014). Indexing the approximate number system. *Acta Psychologica, 145*, 147–155. doi: 10.1016/j.actpsy.2013.11.009

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*(8), e124. doi: 10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2008). Why Most Discovered True Associations Are Inflated. *Epidemiology*, *19*(5), 640–648. doi: 10.1097/EDE.0b013e31818131e7

Ioannidis, J. P. A. (2010). Meta-research: The art of getting it wrong. *Research Synthesis Methods*, *1*(3–4), 169–184. doi: 10.1002/jrsm.19

Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin & Review*, *9*(4), 637–671. doi: 10.3758/BF03196323

Kolkman, M. E., Kroesbergen, E. H., & Leseman, P. P. M. (2013). Early numerical development and the role of non-symbolic and symbolic skills. *Learning and Instruction*, *25*, 95–103. doi: 10.1016/j.learninstruc.2012.12.001

Krajewski, K., & Schneider, W. (2009). Exploring the impact of phonological awareness, visual-spatial working memory, and preschool quantity-number competencies on mathematics achievement in elementary school: Findings from a 3-year longitudinal study. *Journal of Experimental Child Psychology*, *103*(4), 516–531. doi: 10.1016/j.jecp.2009.03.009

Leibovich, T., & Henik, A. (2013). Magnitude processing in non-symbolic stimuli. *Frontiers in Psychology*, *4*, 375. doi. 10.3389/fpsyg.2013.00375

Leibovich, T., Katzin, N., Harel, M., & Henik, A. (2017). From "sense of number" to "sense of magnitude": The role of continuous magnitudes in numerical cognition. *Behavioral and Brain Sciences*, *40*, e164. doi: 10.1017/S0140525X16000960

Li, Y., & Geary, D. C. (2013). Developmental Gains in Visuospatial Memory Predict Gains in Mathematics Achievement. *PLoS ONE*, *8*(7). doi: 10.1371/journal.pone.0070160

Libertus, M. E., Feigenson, L., & Halberda, J. (2011). Preschool acuity of the approximate number system correlates with school math ability. *Developmental Science*, *14*(6), 1292–1300. doi: 10.1111/j.1467-7687.2011.01080.x

Libertus, M. E., Feigenson, L., & Halberda, J. (2013). Is approximate number precision a stable predictor of math ability? *Learning and Individual Differences*, *25*, 126–133. doi: 10.1016/J.LINDIF.2013.02.001

Logie, R. H., & Baddeley, A. D. (1987). Cognitive Processes in Counting. *Journal of Experimental Psychology. Learning, Memory, and Cognition.*, *13*(2), 310–326.

Lyons, I. M., Bugden, S., Zheng, S., De Jesus, S., & Ansari, D. (2018). Symbolic number skills predict growth in nonsymbolic number skills in kindergarteners. *Developmental Psychology, 54*(3), 440. doi: 10.1037/dev0000445

Lyons, I. M., Nuerk, H. C., & Ansari, D. (2015). Rethinking the implications of numerical ratio effects for understanding the development of representational precision and numerical processing across formats. *Journal of Experimental Psychology: General*, *144*(5), 1021–1035. doi: 10.1037/xge0000094

Lyons, I. M., Price, G. R., Vaessen, A., Blomert, L., & Ansari, D. (2014). Numerical predictors of arithmetic success in grades 1-6. *Developmental Science*, *17*(5), 714–726. doi: 10.1111/desc.12152

Maindonald, J. H. and Braun, W. J. (2015). *DAAG: Data Analysis and Graphics Data and Functions*. R package version 1.22.

Mammarella, I. C., Caviola, S., Cornoldi, C., & Lucangeli, D. (2013). Mental additions and verbal-domain interference in children with developmental dyscalculia. *Research in Developmental Disabilities*, *34*(9), 2845–2855. doi: 10.1016/j.ridd.2013.05.044

Mammarella, I. C., Caviola, S., Giofrè, D., & Szűcs, D. (2018). The underlying structure of visuospatial working memory in children with mathematical learning disability. *British Journal of Developmental Psychology*, *36*(2), 220–235. doi: 10.1111/bjdp.12202

McCabe, D. P., Roediger, H. L., McDaniel, M. A., Balota, D. A., Hambrick, D. Z., & Hambrick, D. Z. (2010). The relationship between working memory capacity and executive functioning: evidence for a common executive attention construct. *Neuropsychology*, *24*(2), 222–243. doi: 10.1037/a0017619

Meehl, P. E. (1967). Theory-Testing in Psychology and Physics: A Methodological Paradox. *Philosophy of Science*, *34*(2), 103–115. doi: 10.1086/288135

Melby-Lervåg, M., Redick, T. S., & Hulme, C. (2016). Working Memory Training Does Not Improve Performance on Measures of Intelligence or Other Measures of "Far Transfer." *Perspectives on Psychological Science*, *11*(4), 512–534. doi: 10.1177/1745691616635612

Menon, V. (2016). Working memory in children's math learning and its disruption in dyscalculia. *Current Opinion in Behavioral Sciences*, *10*, 125–132. doi: 10.1016/J.COBEHA.2016.05.014

Meyer, M. L., Salimpoor, V. N., Wu, S. S., Geary, D. C., & Menon, V. (2010). Differential contribution of specific working memory components to mathematics achievement in 2nd and 3rd graders. *Learning and Individual Differences*, *20*(2), 101–109. doi: 10.1016/j.lindif.2009.08.004

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature*, *215*(5109), 1519–1520. doi: 10.1038/2151519a0

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. doi: 10.1371/journal.pone.0067918

Olivola, C. Y., & Chater, N. (2017). Numerical magnitude evaluation as a foundation for decision making. *Behavioral and Brain Sciences, 40*, e183. doi: 10.1017/S0140525X16002211

Passolunghi, M. C., & Mammarella, I. C. (2010). Spatial and visual working memory ability in children with difficulties in arithmetic word problem solving. *European Journal of Cognitive Psychology*, *22*(6), 944–963. doi: 10.1080/09541440903091127

Passolunghi, M. C., & Mammarella, I. C. (2012). Selective Spatial Working Memory Impairment in a Group of Children With Mathematics Learning Disabilities and Poor Problem-Solving Skills. *Journal of Learning Disabilities*, *45*(4), 341–350. doi: 10.1177/0022219411400746

Passolunghi, M. C., Mammarella, I. C., & Altoè, G. (2008). Cognitive Abilities as Precursors of the Early Acquisition of Mathematical Skills During First Through Second Grades. *Developmental Neuropsychology*, *33*(3), 229–250. doi: 10.1080/87565640801982320

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), 455–473. doi: 10.1037/edu0000079

Peng, P., Wang, C., & Namkung, J. (2018). Understanding the Cognition Related to Mathematics Difficulties: A Meta-Analysis on the Cognitive Deficit Profiles and the Bottleneck Theory. *Review of Educational Research*, *88*(3), 434–476. doi: 10.3102/0034654317753350

Price, G. R., Palmer, D., Battista, C., & Ansari, D. (2012). Nonsymbolic numerical magnitude comparison: Reliability and validity of different task variants and outcome measures, and their relationship to arithmetic achievement in adults. *Acta Psychologica*, *140*(1), 50–57. doi: 10.1016/j.actpsy.2012.02.008

Price, G. R., & Wilkey, E. D. (2017). Cognitive mechanisms underlying the relation between nonsymbolic and symbolic magnitude processing and their relation to math. *Cognitive Development, 44*, 139–149. doi: 10.1016/j.cogdev.2017.09.003

Raghubar, K. P., Barnes, M. A., & Hecht, S. A. (2010). Working memory and mathematics: A review of developmental, individual difference, and cognitive approaches. *Learning and Individual Differences*, *20*(2), 110–122. doi: 10.1016/J.LINDIF.2009.10.005

Rasmussen, C., & Bisanz, J. (2005). Representation and working memory in early arithmetic. *Journal of Experimental Child Psychology*, *91*(2), 137–157. doi: 10.1016/J.JECP.2005.01.004

Sartori, G., Job, R., & Tressoldi, P. E. (1995). *Battery for the assessment of developmental dyslexia and dysorthographia* [Batteria per la valutazione della dislessia e della disortografia evolutiva]. Florence: Giunti OS.

Sasanguie, D., De Smedt, B., Defever, E., & Reynvoet, B. (2012). Association between basic numerical abilities and mathematics achievement. *British Journal of Developmental Psychology*, *30*(2), 344–357. doi: 10.1111/j.2044-835X.2011.02048.x

Sasanguie, D., de Smedt, B., & Reynvoet, B. (2015, January 26). Evidence for distinct magnitude systems for symbolic and non-symbolic number. *Psychological Research*, pp. 1–12. Springer Berlin Heidelberg. doi: 10.1007/s00426-015-0734-1

Sasanguie, D., Defever, E., Maertens, B., & Reynvoet, B. (2014). The approximate number system is not predictive for symbolic number processing in kindergarteners. *The Quarterly Journal of Experimental Psychology*, *67*(2), 271–280. doi: 10.1080/17470218.2013.803581

Sasanguie, D., Göbel, S. M., Moll, K., Smets, K., & Reynvoet, B. (2013). Approximate number sense, symbolic number processing, or number–space mappings: What underlies mathematics achievement? *Journal of Experimental Child Psychology*, *114*(3), 418–431. doi: 10.1016/j.jecp.2012.10.012

Schneider, M., Beeres, K., Coban, L., Merz, S., Susan Schmidt, S., Stricker, J., & De Smedt, B. (2017). Associations of non-symbolic and symbolic numerical magnitude processing with mathematical competence: a meta-analysis. *Developmental Science*, n/a-n/a. doi: 10.1111/desc.12372

Schwenk, C., Sasanguie, D., Kuhn, J. T., Kempe, S., Doebler, P., & Holling, H. (2017). (Non-)symbolic magnitude processing in children with mathematical difficulties: a meta-analysis. *Research in Developmental Disabilities*, *64*, 152–167. doi: 10.1016/j.ridd.2017.03.003

Smets, K., Sasanguie, D., Szücs, D., & Reynvoet, B. (2015). The effect of different methods to construct non-symbolic stimuli in numerosity estimation and comparison. *Journal of Cognitive Psychology*, *27*(3), 310–325. doi: 10.1080/20445911.2014.996568

Swanson, H. L., & Sachse-Lee, C. (2001). Mathematical Problem Solving and Working Memory in Children with Learning Disabilities: Both Executive and Phonological Processes Are Important. *Journal of Experimental Child Psychology*, *79*(3), 294–321. doi: 10.1006/JECP.2000.2587

Szűcs, D. (2016). Chapter 11 – Subtypes and comorbidity in mathematical learning disabilities: Multidimensional study of verbal and visual memory processes is key to understanding. In *Progress in Brain Research* (Vol. 227, pp. 277–304). doi: 10.1016/bs.pbr.2016.04.027

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

Szűcs, D., Devine, A., Soltesz, F., Nobes, A., & Gabriel, F. (2014). Cognitive components of a mathematical processing network in 9-year-old children. *Developmental Science*, *17*(4), 506–524. doi: 10.1111/desc.12144

Szűcs, D., & Ioannidis, J. P. A. (2017a). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, *15*(3). doi: 10.1371/journal.pbio.2000797

Szűcs, D., & Ioannidis, J. P. A. (2017b). When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment. *Frontiers in Human Neuroscience*, *11*, 390. doi: 10.3389/fnhum.2017.00390

Szűcs, D., & Myers, T. (2017). A critical analysis of design, facts, bias and inference in the approximate number system training literature: A systematic review. *Trends in Neuroscience and Education*, *6*, 187–203. doi: 10.1016/j.tine.2016.11.002

Szűcs, D., Nobes, A., Devine, A., Gabriel, F. C., & Gebuis, T. (2013b). Visual stimulus parameters seriously compromise the measurement of approximate number system acuity and comparative effects between adults and children. *Frontiers in Psychology*, *4*, 444. doi: 10.3389/fpsyg.2013.00444

Szũcs, D., Soltész, F., Jármi, É., & Csépe, V. (2007). The speed of magnitude processing and executive functions in controlled and automatic number comparison in children: an electro-encephalography study. *Behavioral and Brain Functions*, *3*(3). doi: 10.1186/1744-9081-3-23

Swanson, H. L. (2017). Verbal and visual-spatial working memory: What develops over a life span? *Developmental Psychology, 53*(5), 971. doi: 10.1037/dev0000291

Tokita, M., & Ishiguchi, A. (2013). Effects of perceptual variables on numerosity comparison in 5–6-year-olds and adults. *Frontiers in Psychology*, *4*, 431. doi: 10.3389/fpsyg.2013.00431

Tosto, M. G., Petrill, S. A., Malykh, S., Malki, K., Haworth, C. M. A., Mazzocco, M. M. M., … Kovas, Y. (2017). Number Sense and Mathematics: Which, When and How? *Developmental Psychology*, (August). doi: 10.1037/dev0000331

Van Opstal, F., Gevers, W., De Moor, W., & Verguts, T. (2008). Dissecting the symbolic distance effect: Comparison and priming effects in numerical and nonnumerical orders. *Psychonomic Bulletin & Review*, *15*(2), 419–425. doi: 10.3758/PBR.15.2.419

Vanbinst, K., Ansari, D., Ghesquière, P., De Smedt, B., Ansari, D., & Ross, J. (2016). Symbolic numerical magnitude processing is as important to arithmetic as phonological awareness is to reading. *PLOS ONE*, *11*(3), e0151045. doi: 10.1371/journal.pone.0151045

vanMarle, K., Chu, F. W., Li, Y., & Geary, D. C. (2014). Acuity of the approximate number system and preschoolers' quantitative development. *Developmental Science, 17*, 492–505. doi: 10.1111/desc.12143

Wagenmakers, E.-J., Verhagen, J., & Ly, A. (2016). How to quantify the evidence for the absence of a correlation. *Behavior Research Methods*, *48*(2), 413–426. doi: 10.3758/s13428-015-0593-0

Wei, W., Lu, H., Zhao, H., Chen, C., Dong, Q., & Zhou, X. (2012). Gender Differences in Children's Arithmetic Performance Are Accounted for by Gender Differences in Language Abilities. *Psychological Science*, *23*(3), 320–330. doi: 10.1177/0956797611427168

Wong, T. T.-Y., Ho, C. S.-H., & Tang, J. (2016). The relation between ANS and symbolic arithmetic skills: The mediating role of number–numerosity mappings. Contemporary Educational Psychology, 46, 208–217. doi: 10.1016/j.cedpsych.2016.06.003

Xenidou-Dervou, I., De Smedt, B., van der Schoot, M., & van Lieshout, E. C. D. M. (2013). Individual differences in kindergarten math achievement: The integrative roles of approximation skills and working memory. Learning and Individual Differences, 28, 119–129. doi: 10.1016/j.lindif.2013.09.012

Xenidou-Dervou, I., Molenaar, D., Ansari, D., van der Schoot, M., & van Lieshout, E. C. D. M. (2017). Nonsymbolic and symbolic magnitude comparison skills as longitudinal predictors of mathematical achievement. *Learning and Instruction*, *50*, 1–13. doi: 10.1016/J.LEARNINSTRUC.2016.11.001

Xenidou-Dervou, I., Van Luit, J. E. H., Kroesbergen, E. H., Friso-van den Bos, I., Jonkman, L. M., van der Schoot, M., & van Lieshout, E. C. D. M. (2018). Cognitive predictors of children's development in mathematics achievement: A latent growth modeling approach. *Developmental Science*, *21*(6), e12671. doi: 10.1111/desc.12671

Table 1. *Schematic description of the four versions (Version A, Version B, Version C, and Version D) of the four regression models. Version B–C contain subsets of the predictors included in Version A.*

Model Name	**Version A**	**Version B**	**Version C**	**Version D**
Model Description	Full model	Model A excluding ratio and distance effects	Model B excluding all SNC and ANS measures	Model A include STM and WM measures
Predictors	ANS measures[1]	—	—	ANS measures[1]
	SNC distance effects	—	—	SNC distance effects
	SNC RT/Accuracy	SNC RT/Accuracy	—	SNC RT/Accuracy
	STM measures (Verbal/Spatial)	STM measures (Verbal/Spatial)	STM measures (Verbal/Spatial)	—
	WM measures (Verbal/Spatial)	WM measures (Verbal/Spatial)	WM measures (Verbal/Spatial)	—

[1] The ANS measures included differ between the Model 1 (weber fraction), Model 2 (weber fraction), Model 3 (ANS accuracy), and Model 4 (ANS accuracy).

Table 2. *Overall sample (N), demographic information and descriptive statistics (means and standard errors) of the achievement and intelligence measures for each grade are shown.*

Variables	2nd Grade	4th Grade	6th Grade
Demographics			
Overall sample (N)	413	391	450
Gender: Male; Female	206; 207	197; 194	245; 205
Age in moths (range)	94 (86–106)	119 (109–136)	144 (129–163)
Achievement tasks (z-score)			
Maths composite score (SE)	0.03 (0.05	0.02 (0.05)	0.02 (0.05)
Reading rate composite score (SE)	0.15 (0.05	0.34 (0.05)	-0.08 (0.04)
Intelligence measure			
Cattell (SE)	22.44. (0.30)	28.20 (0.27)	30.64 (0.24)

Table 3. *Descriptive statistics of the magnitude comparison and working memory measures are reported for each grade. The number of observations (N), means (M) and standard errors (SE) are shown for each variable. Because w can only be estimated when accuracy is above 55% it is not possible to estimate w for all participants and therefore the number of cases differ for the different weber fraction estimates.*

Variables		2nd Grade		4th Grade		6th Grade
ANS measures	*N*	*M (SE)*	*N*	*M (SE)*	*N*	*M (SE)*
weber fraction	376	0.52 (0.02)	376	0.40 (0.01)	442	0.38 (0.01)
weber fraction (congruent trials)	381	0.35 (0.02)	376	0.22 (0.01)	443	0.20 (0.01)
weber fraction (incongruent trials)	284	0.64 (0.02)	308	0.55 (0.02)	371	0.52 (0.01)
ANS accuracy	413	0.69 (0.005)	391	0.73 (0.004)	450	0.74 (0.004)
ANS accuracy (congruent trials)	413	0.77 (0.01)	391	0.83 (0.01)	450	0.85 (0.005)
ANS accuracy (incongruent trials)	413	0.60 (0.005)	391	0.63 (0.005)	450	0.64 (0.005)
ANS RT (ms)	413	1480 (26)	391	1306 (22)	450	1173 (20)
ANS RT (congruent trials)	413	1426 (23)	391	1246 (20)	450	1109 (18)
ANS RT (incongruent trial)	413	1559 (32)	391	1389 (26)	450	1261 (24)
SNC measures						
SNC accuracy	413	0.93 (0.004)	391	0.96 (0.003)	450	0.97 (0.002)
SNC distance effect (accuracy)	413	0.06 (0.004)	391	0.04 (0.003)	450	0.03 (0.002)
SNC RT	413	995 (9.91)	391	811 (8.60)	450	696 (7.04)
SNC distance effect (RT)	413	-94 (5.12)	391	-85 (3.85)	450	-59 (2.71)
Working memory tasks						

verbal WM	413	0.41 (0.005)	391	0.52 (0.006)	450	0.60 (0.006)
spatial WM	413	0.49 (0.009)	391	0.64 (0.007)	450	0.71 (0.006)
verbal STM	413	0.63 (0.003)	391	0.68 (0.003)	450	0.70 (0.002)
spatial STM	413	0.80 (0.004)	391	0.86 (0.003)	450	0.89 (0.002)

Table 4. *Summary of zero order and partial correlations with Mathematics composite score. Partial correlations considered the effect of fluid intelligence. There are three columns for each grade. The first value in the 'r' column shows the zero order correlation, the second value shows the partial correlation. The 'zero' and 'partial' columns detail Bayesian inference results for zero order and partial correlations, respectively. The columns show whether the null or alternative hypotheses were supported and the largeness of the Bayes Factors is also indicated. The first number indicates whether the null (0) or the alternative (1) hypothesis was supported. The second number following a + or – sign indicates the largeness of the Bayes Factor. The larger is the absolute value of the number, the stronger is the evidence (0=weak ; 1=substantial ; 2=strong ; 3=very strong ; 4=decisive). In order to facilitate reading the table the second number is negative if the null hypothesis was supported and the second number is positive if the alternative hypothesis was supported. For example, '1+4' means that the alternative hypothesis was supported and the evidence was decisive. '0-2' means that the null hypothesis was supported and the evidence was strong. Additionally, correlations marked with an asterisk indicate that once fluid intelligence was controlled for through partial correlation the evidence switched from being in favour of a correlation to being in favour of the null.*

	Grade 2	Grade 4	Grade 6

Measure		*r*		zero	partial		*r*		zero	partial		*r*		zero	partial
weber fraction	-0.13	-0.10	*	1+0	0-0	-0.13	-0.10	*	1+0	0-0	-0.18	-0.07	*	1+4	0-1
weber fraction (congruent trials)	-0.13	-0.11	*	1+1	0-0	-0.01	0.01		0-2	0-2	-0.18	-0.1	*	1+3	0-0
weber fraction (incongruent trials)	-0.06	-0.04		0-1	0-2	-0.10	-0.09		0-1	0-1	-0.06	-0.02		0-1	0-2
ANS accuracy	0.15	0.09	*	1+1	0-0	0.17	0.13		1+2	1+0	0.16	0.06	*	1+2	0-1
ANS accuracy (congruent trials)	0.12	0.09	*	1+0	0-0	0.08	0.05		0-1	0-1	0.18	0.11	*	1+4	0-0
ANS accuracy (incongruent trials)	0.11	0.05		0-0	0-1	0.20	0.17		1+4	1+2	0.08	-0.02		0-1	0-2
ANS RT	-0.07	-0.06		0-1	0-1	-0.03	0.00		0-2	0-2	-0.12	-0.09	*	1+0	0-0
ANS RT (congruent trials)	-0.08	-0.06		0-1	0-1	-0.05	-0.01		0-2	0-2	-0.14	-0.12		1+1	1+0
ANS RT (incongruent trials)	-0.06	-0.06		0-1	0-1	-0.02	0.00		0-2	0-2	-0.1	-0.07		0-0	0-1
SNC accuracy	0.29	0.22		1+4	1+4	0.16	0.14		1+2	1+0	0.16	0.13		1+2	1+0
SNC distance effect (accuracy)	-0.11	-0.10		0-0	0-0	-0.12	-0.11	*	1+0	0-0	0.01	0.03		0-2	0-2
SNC RT	-0.21	-0.14		1+4	1+1	-0.17	-0.11	*	1+2	0-0	-0.37	-0.26		1+4	1+4
SNC distance effect (RT)	0.00	-0.03		0-2	0-2	0.14	0.12		1+0	1+0	0.15	0.11	*	1+1	0-0
spatial STM	0.25	0.17		1+4	1+3	0.13	0.06	*	1+0	0-1	0.23	0.11	*	1+4	0-0
verbal STM	0.29	0.22		1+4	1+4	0.31	0.25		1+4	1+4	0.33	0.31		1+4	1+4
spatial WM	0.32	0.22		1+4	1+4	0.26	0.17		1+4	1+2	0.41	0.24		1+4	1+4
verbal WM	0.33	0.25		1+4	1+4	0.27	0.21		1+4	1+4	0.34	0.23		1+4	1+4
Cattell (IQ)	0.38	—		1+4	—	0.36	—	*	1+4	—	0.49	—	*	1+4	—
reading composite score (errors)	-0.30	-0.24		1+4	1+4	-0.33	-0.29		1+4	1+4	-0.32	-0.24		1+4	1+4

reading composite score (rate)	0.37	0.34	1+4	1+4	0.38	0.36	1+4	1+4	0.4	0.36	1+4	1+4

Table 5. *Summary of zero order and partial correlations with Reading decoding composite score. Partial correlations considered the effect of fluid intelligence. There are three columns for each grade. The first value in the 'r' column shows the zero order correlation, the second value shows the partial correlation. The 'zero' and 'partial' columns detail Bayesian inference results for zero order and partial correlations, respectively. The columns show whether the null or alternative hypotheses were supported and the largeness of the Bayes Factors is also indicated. The first number indicates whether the null (0) or the alternative (1) hypothesis was supported. The second number following a + or – sign indicates the largeness of the Bayes Factor. The larger is the absolute value of the number, the stronger is the evidence (0=weak ; 1=substantial ; 2=strong ; 3=very strong ; 4=decisive). In order to facilitate reading the table the second number is negative if the null hypothesis was supported and the second number is positive if the alternative hypothesis was supported. For example, '1+4' means that the alternative hypothesis was supported and the evidence was decisive. '0-2' means that the null hypothesis was supported and the evidence was strong. Additionally, correlations marked with an asterisk indicate that once fluid intelligence was controlled for through partial correlation the evidence switched from being in favour of a correlation to being in favour of the null.*

	Grade 2		Grade 4		Grade 6	

Measure		*r*		zero	partial		*r*		zero	partial		*r*		zero	partial
weber fraction	-0.03	-0.01		0-2	0-2	-0.05	-0.04		0-1	0-2	-0.14	-0.09	*	1+1	0-0
weber fraction (congruent trials)	-0.02	-0.01		0+2	0-2	0.02	0.03		0-2	0-2	-0.12	-0.09	*	1+0	0-0
weber fraction (incongruent trials)	-0.07	-0.06		0+1	0-1	-0.01	0.00		0-2	0-2	-0.05	-0.04		0-2	0-2
ANS accuracy	-0.01	-0.04		0+2	0-2	0.04	0.01		0-2	0-2	0.10	0.06		0-0	0-1
ANS accuracy (congruent trials)	0.00	-0.01		0+2	0-2	0.00	-0.02		0-2	0-2	0.15	0.12		1+2	1+0
ANS accuracy (incongruent trials)	-0.03	-0.06		0+2	0-1	0.07	0.05		0-1	0-2	0.01	-0.02		0-2	0-2
ANS RT	-0.18	-0.18		1+3	1+3	-0.07	-0.06		0-1	0-1	-0.05	-0.04		0-1	0-2
ANS RT (congruent trials)	-0.17	-0.16		1+2	1+2	-0.07	-0.05		0-1	0-1	-0.08	-0.06		0-1	0-1
ANS RT (incongruent trials)	-0.18	-0.18		1+3	1+3	-0.08	-0.08		0-1	0-1	-0.03	-0.02		0-2	0-2
SNC accuracy	0.13	0.09	*	1+0	0-1	0.11	0.10		0-0	0-0	0.08	0.06		0-1	0-1
SNC distance effect (accuracy)	-0.02	-0.01		0-2	0-2	-0.11	-0.10		0-0	0-0	-0.02	-0.01		0-2	0-2
SNC RT	-0.14	-0.11	*	1+1	0-0	-0.19	-0.17		1+3	1+2	-0.21	-0.17		1+4	1+3
SNC distance effect (RT)	-0.03	-0.05		0-2	0-2	0.10	0.09		0-0	0-1	0.09	0.07		0-0	0-1
spatial STM	0.17	0.13		1+2	1+0	0.06	0.03		0-1	0-2	0.10	0.05		0-0	0-1
verbal STM	0.16	0.12		1+2	1+0	0.26	0.23		1+4	1+4	0.27	0.25		1+4	1+4
spatial WM	0.22	0.17		1+4	1+3	0.13	0.09	*	1+0	0-1	0.15	0.08	*	1+2	0-1
verbal WM	0.24	0.20		1+4	1+4	0.28	0.25		1+4	1+4	0.28	0.24		1+4	1+4
Cattell (IQ)	0.17	—		1+2	—	0.15	—		1+1	—	0.19	—		1+4	—
reading composite score (errors)	-0.43	-0.41		1+4	1+4	-0.48	-0.46		1+4	1+4	-0.50	-0.48		1+4	1+4

mathematics composite score	0.37	0.34	1+4	1+4	0.38	0.36	1+4	1+4	0.40	0.36	1+4	1+4

Table 6. *Summary of the regression models showing the preferred version and R^2 for each model, outcome measure, and Grade.*

Model	Outcome	Grade	Preferred Version	R^2
Model 1	Math	2	Version B	0.220
		4	Version B	0.186
		6	Version B	0.321
	Reading	2	Version C	0.080
		4	Version B	0.124
		6	Version B	0.137
Model 2	Math	2	Version A	0.315
		4	Version C	0.141
		6	Version B	0.269
	Reading	2	Version B	0.090
		4	Version B	0.117
		6	Version B	0.117
Model 3	Math	2	Version B	0.239
		4	Version B	0.186
		6	Version A	0.337
	Reading	2	Version C	0.094
		4	Version B	0.132
		6	Version B	0.137
Model 4	Math	2	Version B	0.239
		4	Version A	0.205
		6	Version A	0.340
	Reading	2	Version C	0.094
		4	Verison B	0.132
		6	Version B	0.137

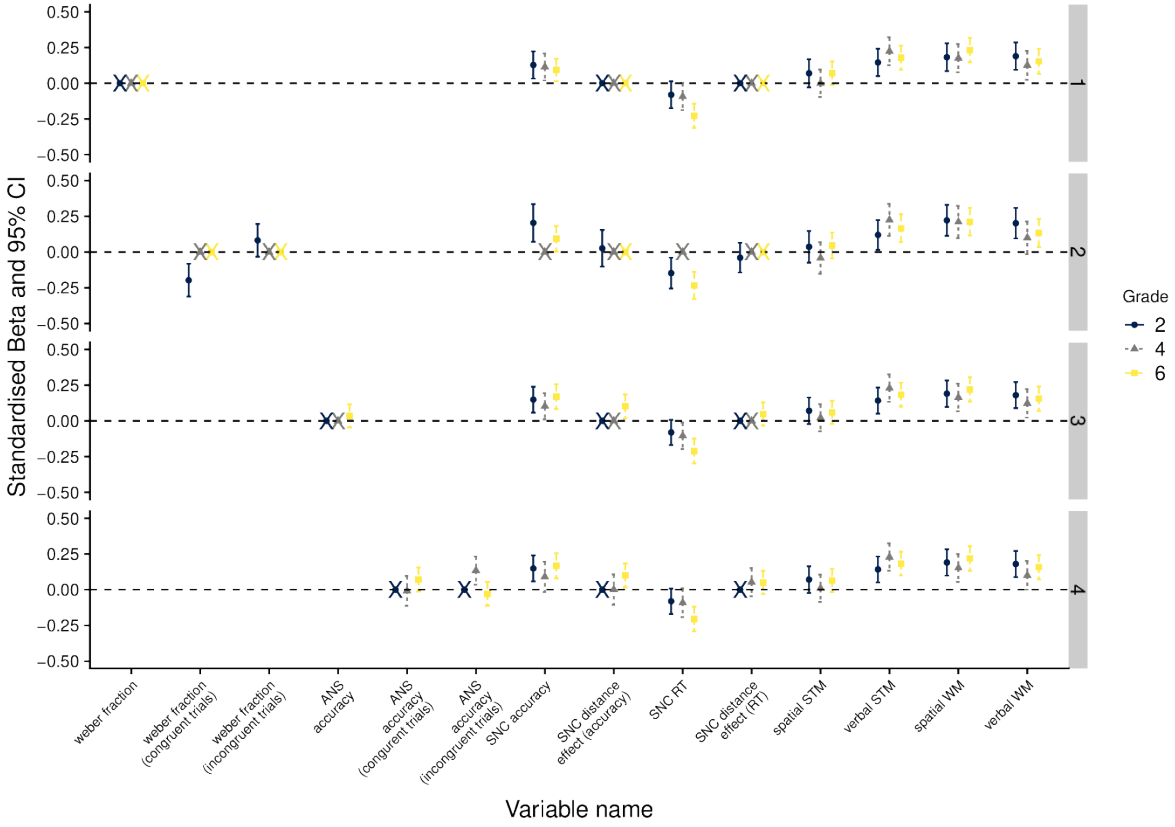

Figure 1. Standardized betas (and 95% CI confidence intervals) for each of the four specifications of the mathematics model. Predictors that were contained in the full model, but dropped from the preferred model, are marked with an X.

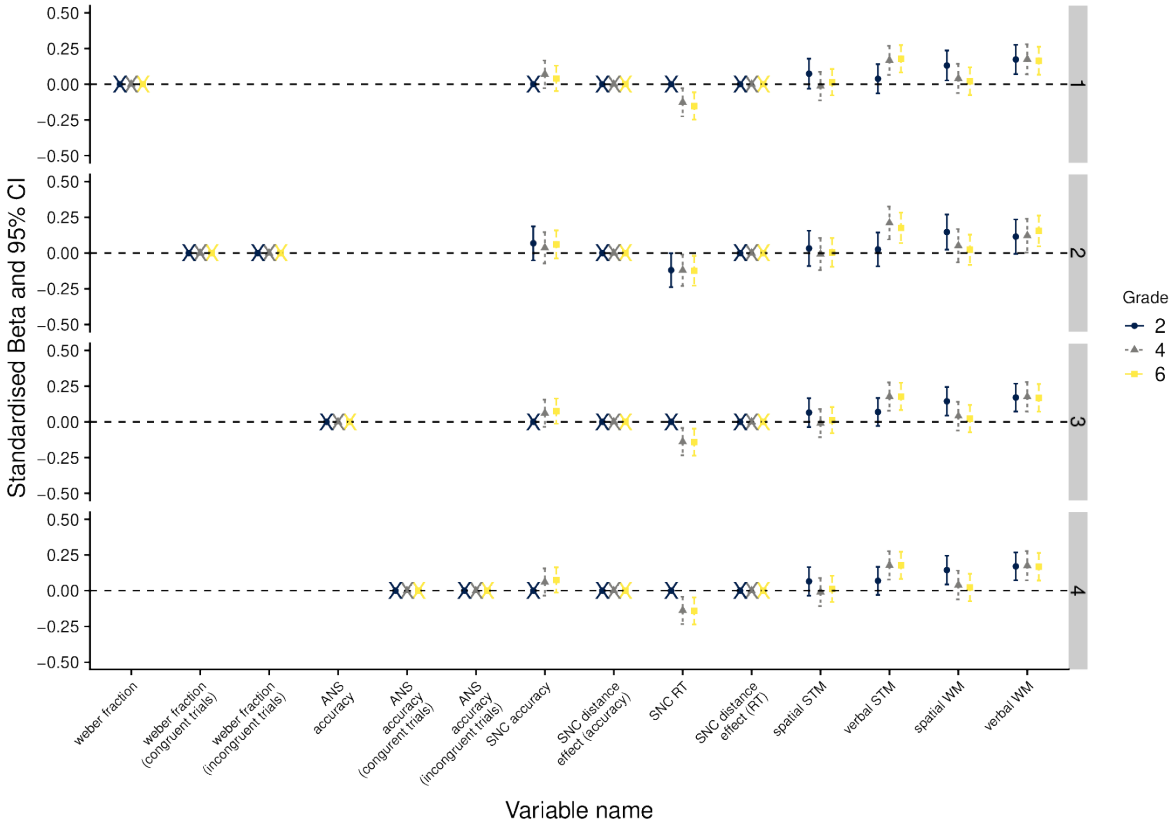

Figure 2. Standardized betas (and 95% CI confidence intervals) for each of the four specifications of the reading model. Predictors that were contained in the full model, but dropped from the preferred model, are marked with an X.