Making Sense of the Relation between Number Sense and Math

Bert Reynvoet, Andrew D. Ribner, Leanne Elliott, Manon Van Steenkiste, Delphine Sasanguie, Melissa E. Libertus

This paper is accepted in Journal of Numerical Cognition

Author Notes

Bert Reynvoet, Manon Van Steenkiste, and Delphine Sasanguie: Brain and Cognition, KU Leuven, Leuven, Belgium & Faculty of Psychology and Educational Sciences, KU Leuven @Kulak, Kortrijk, Belgium.

Delphine Sasanguie, HOGENT, University College Ghent, Orthopedagogy - Special Education Department, Ghent, Belgium.

Andrew D. Ribner, Leanne Elliott, and Melissa E. Libertus: Department of Psychology, University of Pittsburgh, US.

ORCiD

Bert Reynvoet: 0000-0002-4898-2475

Andrew D. Ribner: 0000-0001-7188-223X

Leanne Elliott: 0000-0001-7392-2856

Delphine Sasanguie: 000-0002-3356-5097

Melissa E. Libertus : 0000-0002-6461-1156

Correspondence concerning this article should be addressed to Bert Reynvoet, Faculty of Psychology and Educational Sciences, KU Leuven @Kulak, Etienne Sabbelaan 51, 8500 Kortrijk, Belgium. email: bert.reynvoet@kuleuven.be

Abstract

While several studies have shown that the performance on numerosity comparison tasks is related to individual differences in math abilities, others have failed to find such a link. These inconsistencies could be due to variations in which math was assessed, different stimulus generation protocols for the numerosity comparison task, or differences in inhibitory control. This within-subject study is a conceptual replication tapping into the relation between numerosity comparison, math, and inhibition in adults (N=122). Three aspects of math ability were measured using standardized assessments: Arithmetic fluency, calculation, and applied problem solving skills. Participants' inhibitory skills were measured using Stroop and Go/No-Go tasks with numerical and non-numerical stimuli. Finally, non-symbolic number sense was measured using two different versions of a numerosity comparison task that differed in the stimulus generation protocols (Panamath, Halberda et al., 2008; G&R, Gebuis & Reynvoet, 2011). We find that performance on the Panamath task, but not the G&R task, related to measures of calculation and applied problem solving but not arithmetic fluency, even when controlling for inhibitory control. One possible explanation is that depending on the characteristics of the stimuli in the numerosity comparison task, the reliance on numerical and non-numerical information may vary and only when performance relies more on numerical representations, a relation with math achievement is found. Our findings help to explain prior mixed findings regarding the link between non-symbolic number sense and math and highlight the need to carefully consider variations in numerosity comparison tasks and math measures.

Keywords: numerosity processing, inhibition, mathematics performance, number sense

Introduction

Humans are born with an evolutionarily preserved intuitive sense of approximate, non-symbolic number, commonly referred to as number sense (Dehaene, 2001)—that helps us to make important decisions without the need for exact enumeration. This idea is supported by findings that animals (Agrillo, Miletto Petrazzini, & Bisazza 2016; Viswanathan & Nieder, 2015) and infants (Libertus & Brannon, 2010; Starr, Libertus, & Brannon, 2013; Xu, Spelke, & Goddard, 2005) can discriminate numerosities. In older children and adults, number sense is typically assessed with a numerosity comparison task in which participants are instructed to decide which of two presented numerosities is numerically larger (i.e., contains the larger number of objects). Performance on numerosity comparison tasks is characterized by a ratio effect: when the relative difference between both numerosities is small, decisions are harder and participants are less accurate than when the relative differences are larger (e.g., Halberda, Ly, Wilmer, Naiman, & Germine, 2012; Libertus, Odic, & Halberda, 2012; Sasanguie, Göbel, Moll, Smets, & Reynvoet, 2013). Performance on numerosity comparison tasks improves throughout development into adulthood: older children can progressively discriminate smaller differences, and this ability continues to develop into adulthood when people can discriminate, on average, an 8:9 ratio (Halberda & Feigenson, 2008; Defever, Reynvoet, & Gebuis, 2013; Piazza, De Feo, Panzeri, & Dehaene, 2018).

In addition to this non-symbolic number sense, humans have a unique capacity to represent and perform operations on symbolic numbers that form the basis of more advanced mathematics, and some have suggested that the understanding of symbolic number is built upon non-symbolic number sense. In the first study to suggest this association, Halberda and colleagues (2008) demonstrated that performance on a numerosity comparison task in adolescence was retrospectively related to performance on a symbolic mathematics achievement test as early as kindergarten. Since then, the relation between performance on numerosity comparison tasks and age-appropriate measures of mathematics achievement has been replicated in several cross-sectional and longitudinal studies with children and adults (e.g., Halberda et al., 2012; Libertus et al., 2012; Libertus, Feigenson, & Halberda, 2011; Nys & Content, 2012; Piazza, Pica, Izard, Spelke, & Dehaene, 2013; Starr, Libertus, & Brannon, 2013), suggesting that non-symbolic number processing may be the basis of symbolic mathematics. In addition, children with

dyscalculia also show lower levels of performance on numerosity comparison tasks (Mazzocco, Feigenson, & Halberda, 2011; Piazza et al., 2010). Finally, further support in favour of an association between numerosity comparison and mathematics achievement comes from intervention studies demonstrating that training numerosity comparison results in significant gains on mathematics achievement tests in both children and adults (Libertus, Odic, Feigenson & Halberda, 2020; Park & Brannon, 2013, 2014; Park, Bermudez, Roberts, & Brannon, 2016).

Despite the abundance of positive results, several other studies have called into question the claim that non-symbolic number sense underlies understanding of symbolic number. A number of studies have reported null associations between performance on numerosity comparison tasks and mathematics achievement (e.g., Sasanguie, De Smedt, & Reynvoet, 2017; Sasanguie, Defever, Maertens, & Reynvoet, 2014; Sasanguie et al., 2013), and given a bias toward publication of significant results, this lack of association might be underrepresented. Meta-analyses in typical samples (Chen & Li, 2014; Fazio, Bailey, Thompson, & Siegler, 2014; Schneider et al., 2017) and in children with mathematical learning deficits (Schwenk, Sasanguie, Kuhn, Kempe, Doebler, & Holling, 2017) suggest a weak to modest relation between numerosity comparison skills and math achievement. More crucially, these meta-analyses describe substantial heterogeneity in the effect sizes across studies, but the sources of this heterogeneity remain unclear.

One reason for these mixed findings may be the different ways that math achievement is measured and thus the underlying math skills included in past studies. Previous studies have indeed used very different indices of math achievement such as age-appropriate mathematical abilities tests (e.g., Halberda et al., 2008; Libertus, Odic, Feigenson, & Halberda, 2016), arithmetic fluency tests (e.g., Brankaer, Ghesquière, & De Smedt, 2014; Sasanguie et al., 2013), or curriculum-based tests (e.g. Sasanguie, De Smedt, Defever, & Reynvoet, 2012). The impact which measure of math is used in these studies was addressed in a meta-analysis by Schneider and colleagues (2017). Specifically, the largest effect sizes were found in studies that tested early mathematical abilities or arithmetic fluency. When mathematics measures focused on written arithmetic or a curriculum-based test, the effect sizes were smaller. However, even for adults the relation between numerosity comparison and math seems to be moderated by the type of math achievement test. Braham and Libertus (2018) showed that the performance in a numerosity

comparison task was related to arithmetic fluency and applied word problem solving, but not to the performance on a procedural calculation test. Thus, it is possible that the relation between numerosity comparison performance and math is stronger when the math achievement test measures abilities that tap into numerosity processing (Schneider et al., 2017).

A second reason for these inconsistent findings regarding the link between numerosity comparison and math achievement may stem from the fact that both skills are related to a third variable, specifically inhibition (Fuhs & McNeil, 2013; Gilmore et al., 2013). This suggestion in part goes back to a related debate in the field on the underlying processes in numerosity comparison. Inhibition is broadly defined as overriding a prepotent or dominant response in favor of one that is less common or readily available. As mentioned above, the common interpretation of performance on a numerosity comparison task is that it reflects an individual's representations of non-symbolic numerical information (Halberda et al., 2008; Piazza et al., 2018). However, visual sets of objects also have non-numerical features such as the total surface and density of all objects that correlate with number in everyday life. For instance, eight oranges in a bowl will have a larger volume (i.e., greater total surface) and will be more compact (i.e., denser) than four oranges in the same bowl. Researchers are well aware of this natural correlation and many solutions have been proposed to eliminate the influence of these non-numerical features in tasks that require numerosity processing by manipulating the extent to which non-numerical cues vary across stimuli. Typically, the non-numerical features are positively correlated with number on half of the trials (i.e., congruent trials), such that the numerically larger set also has larger values on non-numerical dimensions, and negatively on the other half (i.e., incongruent trials), so the numerically larger set actually has smaller values on non-numerical dimensions (see Figure 1). Even after such a manipulation, a numerical ratio effect is present and this is considered evidence that individuals are basing their decisions on numerical information (Halberda et al., 2008; Piazza et al., 2018). However, in addition to a ratio effect, a congruency effect is also observed in most studies: performance on incongruent trials is worse than on congruent trials, indicating interference by non-numerical information (e.g., Clayton, Gilmore, & Inglis, 2015; Defever et al., 2013; Gebuis & Reynvoet, 2012a; Gilmore, Cragg, Hogan, & Inglis, 2016 Fuhs & McNeil, 2013; Fuhs, McNeil, Kelley, O'Rear, & Villano, 2016; Leibovich, & Henik, 2014; Norris, Clayton, Gilmore, Inglis & Castronovo, 2019; Reynvoet, Vos & Henik, 2018; Szűcs, Nobes, Devine, Gabriel, & Gebuis, 2013).

Developmental studies have shown that the interference from non-numerical features is stronger in young children than in older children and adults, possibly due to the fact that their inhibitory functions are not yet as mature (e.g., Jonkman, 2006). In fact, children's performance in a numerosity comparison task is sometimes better predicted by the non-numerical features than by number itself (Rousselle & Noël, 2008). The influence of non-numerical features decreases across development (e.g., Defever et al., 2013) but remains present among adults in some studies (Clayton et al., 2015; Gebuis & Reynvoet, 2012a; Leibovich, & Henik, 2014). This developmental pattern is consistent with developmental changes in performance in conflict tasks like Stroop or Flanker tasks (for an overview see Diamond, 2013), which has resulted in the assumption that increasing performance with age in numerosity comparison reflects—at least in part—individuals' increasing inhibitory control abilities (Defever et al., 2013; Szücs et al., 2013). As such, the relation between the performance in a numerosity comparison task and math achievement may be attributable to individual differences in inhibition, as suggested by Fuhs and McNeil (2013) and Gilmore et al. (2013). More specifically, Gilmore et al. (2013) demonstrated that children's performance on incongruent trials and not the performance on congruent trials was related to mathematics achievement. Furthermore, they found that the relation between numerosity comparison and mathematics achievement was no longer significant when accounting for inhibitory ability, indexed by the difference score between a block where children had to name the shape (circle/square) or direction (up/down) of arrows and an inhibition block in which participants had to give the opposite response, e.g., say circle when a square is presented (i.e., NEPSY-II subtest, Davis & Matthews, 2010). However, other studies failed to find that variations in inhibition entirely explain the association between numerosity comparison and math performance (Keller & Libertus, 2015; Odic, Libertus, Feigenson, Halberda, 2013; Odic, Hock, & Halberda, 2014). For instance, in contrast to Gilmore et al. (2013), Keller and Libertus (2015) found that the same measure of inhibition could not explain the positive association between numerosity comparison and mathematics achievement, at least in children from predominantly middle- and high-SES backgrounds. No studies to date have addressed the relation between numerosity comparison, inhibition, and math in adults.

Finally, the contradictory findings of Gilmore et al. (2013) and Keller and Libertus (2015) open the possibility that there are other reasons for the mixed findings. These studies used two different algorithms to create the visual displays of numerosities shown in the numerosity

comparison tasks (specifically, the task developed by Gebuis & Reynvoet, 2011, versus the "Panamath" task developed by Halberda et al., 2008; hereafter referred to as "G&R" and "Panamath" respectively). Previous studies with mostly adult participants have indeed highlighted some important methodological differences between these two very common algorithms that result in different performance (e.g., Clayton et al., 2015; DeWind & Brannon, 2016; Gilmore, Attridge, de Smedt, & Inglis, 2014; Norris & Castronovo, 2016; Smets, Moors & Reynvoet, 2016; Smets, Sasanguie, & Reynvoet, 2014). Despite purportedly measuring the same underlying construct, a systematic analysis of numerical and non-numerical features by DeWind and Brannon (2016) found within-person performance on G&R and Panamath were weakly correlated. They concluded that the different algorithms to create the dot arrays induce differences in performance and also lead to different levels of reliance on non-numerical features, as the Panamath algorithm induced a larger influence of spacing, a mathematically derived feature related to convex hull and sparsity (i.e., the inverse of density), in the decisions of participants than the G&R algorithm. This may be due to a larger positive correlation between the non-numerical features and number in the Panamath task. In other words, inhibition is less likely to be recruited in order to override salient non-numerical features in the Panamath protocol as in the G&R protocol because those features more frequently correspond to the correct numerical response. As such, the moderating effect of inhibition in the relation between numerosity comparison and math achievement may be dependent on the algorithm used to create the dot arrays.

In this study, we attempt to disentangle seemingly contradictory findings from prior research using a direct and conceptual replication of studies that examined the relation between non-symbolic number sense and symbolic math skills in adults. We opted for an adult sample because we wanted to untangle the complex processes first before exploring how these processes operate through development. This enabled us to examine 1) the possible impact of the type of math abilities that were assessed, 2) the possible role that inhibition may have on the relation between number sense and math, and 3) the stimulus generation algorithm used to construct the dot arrays for the numerosity comparison task in a within-subject design where all participants were presented with a large number of tasks. More concretely, in a sample of 122 adults we compared two non-symbolic number comparison tasks (stimuli were generated with either the Panamath or

the G&R software)[1] and their relations with different inhibition (Stroop and Go/No-Go tasks) and math measures (arithmetic fluency, procedural calculation, and applied problem solving skills). These data allow for the examination of whether numerosity comparison tasks relate to different math skills (arithmetic fluency, procedural calculation, word problem solving) and whether different inhibitory control measures are a confounding third variable in these relations.

Methods

Participants

To determine sample size, an *a priori* power analysis was conducted using G*Power 3.1 (Faul, Erdfelder, Lang, & Buchner, 2007). This analysis showed that a sample of 133 would be needed in order to detect an effect size of $r = 0.24$ (see Schneider et al., 2017) with $\alpha = .05$ and a power of 80%. In total, 141 adult participants recruited from a large university in Belgium (university students; $M_{age} = 20.23$ years; SD = 2.05; range = 17.89-29.40) were tested. Participants were excluded from all analyses if any of their scores on one of the assessments was 3*SD* below or above the sample mean for that assessment ($n = 19$). The resulting sample consisted of 122 adults ($M_{age} = 19.70$ years; $SD = 2.05$). Unfortunately, due to the removal of these outliers, the required sample size derived from our power analysis was not fully met.

Procedure

All assessments took place in a single laboratory visit, and tasks were presented in a pseudo-randomized order. All participants first completed a series of paper-and-pencil questionnaires and math tests (first Arithmetic Fluency, then Arithmetic Skills). Then, both numerosity comparison tasks were administered in a counterbalanced order (half of the participants started with Panamath, the other half with G&R). Finally, four inhibition tasks were administered of which the order was counterbalanced according to a Latin square design. Numerosity comparison and inhibition tasks were computerized and presented on a 15.4-inch laptop computer. In total, testing lasted about one hour.

[1] There are several ways in which the stimuli can be created with each software. The details of the version that was used here can be found in the methods section and the experimental scripts are available on https://osf.io/rvh57/.

Measures

Math tests.

Arithmetic Fluency. To measure arithmetic fluency, the 'Tempo Test Arithmetic' was used ('*Tempo Test Rekenen*', De Vos, 1992). This test is comprised of five separate columns: addition, subtraction, multiplication, division, and mixed exercises with 40 items of increasing complexity in each column (e.g., 1 x 4 as first item of multiplication, 5 x 17 as last item). As this test was developed for elementary school children, we reduced the time limit to 30 seconds per column instead of one minute. Participants were instructed to answer as many items correctly as possible. One point was given for each correct item. This assessment demonstrates adequate psychometric validity and reliability (Ghesquière & Ruijssenaars, 1994).

Arithmetic Skills. Four subtests of the 'Cognitive Skills Arithmetic, 5th grade' test ('*Cognitieve Deelvaardigheden Rekenen – 5e graad*' (Desoete & Roeyers, 2002) were used to measure participants' procedural calculation skills and abilities to solve applied word problems. Developers of the assessment have reported adequate internal consistency for the overall assessment with all nine subtests (α = .75). The measure of ***procedural calculation*** was drawn from two subtests; the *procedural calculation* subtest, which consisted of five items which required participants to complete multi-step operations (e.g., 1263 + 861 + 73 + 445 = ?), and the *mathematical reasoning* subtest, which consisted of five items which required participants to interpret and solve a procedural computation (e.g., 370.5 is 0.9 less than …). The measure of ***applied word problems*** was drawn from two other subtests; the *word problems without distraction* contained 5 items with only item-relevant information (e.g., "Emily has 40 marbles. She gives away 2 marbles. How many marbles does she have left?"), and the *word problems with distraction* contained 5 items with additional item-irrelevant information (e.g., "Lisa has 2 marbles and 3 stickers. She gets 40 marbles. How many marbles does she have now?"). Participants were given 15 minutes time to complete all 20 items.

Inhibitory tasks. Participants completed four computer-based tasks designed to test inhibitory control: Two Stroop tasks, and two Go/No-Go tasks in which one task of each kind used numerically relevant stimuli and the other used numerically irrelevant stimuli. The four tasks were presented in a pseudo-random order counterbalanced across participants.

Stroop tasks. In the Stroop tasks, participants were instructed to respond to one piece of relevant information about a stimulus while ignoring the other, often more salient, piece of information. Stroop tasks consisted of 16 practice trials with feedback[2], followed by 96 test trials. A fixation cross was presented for 500 ms on an otherwise blank screen, after which the stimulus was presented for 90 ms, and participants had 750 ms to respond before the start of the next trial. Half of the trials were congruent, and half were incongruent. During the *Numerical Stroop task* (e.g., Censabella & Noël, 2005), stimuli were digits (1 to 4). Participants were asked to indicate how many numbers they saw on the screen rather than the value of the digit (e.g., if they saw 444, the correct answer was 3). The correct answer ranged from 1 to 4 across trials, with responses counterbalanced so that each was correct for 25% of all trials. Participants could answer with the keys f (for 1), g (for 2), h (for 3) and j (for 4). During the *Animal Stroop task* (e.g., Gilmore et al., 2015) two animal pictures were simultaneously presented in the center of the screen. Participants were asked to indicate which animal is largest in real life rather than the relative size on the screen. During half of the trials the image of the larger real-life animal (elephant or bear) was larger than the smallest animal (butterfly or rabbit) (congruent trials). During the other half of the trials, the image of largest real-life animal was presented smaller than the image of the smallest real-life animal (incongruent trials). The physically larger image was always twice as high and twice as wide as the physically smaller image. There were 16 different stimuli, each repeated 6 times. The correct answer was on the left side of the screen during half of the trials, in the other half on the right side of the screen. Response keys were f (left side) and j (right side). Inhibition was measured by an interference score for each task by subtracting the mean response time (RT) on correct congruent trials from the mean RT on correct incongruent trials.

Go/No-Go tasks. In the Go/No-Go tasks, participants were instructed to respond as quickly as possible to all stimuli except for a specified stimulus, thus requiring them to develop and subsequently inhibit a prepotent response. The task consisted of 10 practice trials followed by 100 test trials, in which the "Go" stimulus was presented for 75% of trials and the "No-Go" stimulus was presented for the remaining 25%. A fixation cross was presented for 500 ms, the stimulus was presented for 90 ms, and participants had 750 ms to respond before the start of the next trial. In the *Number Go/No-Go task*, the "Go" cue was the Arabic numeral "1", while the

[2] 9 participants received only 5 practice trials due to a programming error; however, they were not excluded from analyses.

"No-Go" cue was the Arabic numeral "6". In the *Animal Go/No-Go task*, the "Go" cue was a picture of a horse, while the "No-Go" cue was a picture of a bird. The number of commission errors (i.e., a response was given despite the No-Go stimulus) in each task was used a measure of inhibition.

Numerosity comparison tasks. Participants completed two numerosity comparison tasks for which stimuli were generated with different software (Panamath and G&R). Both tasks started with 6 practice trials with auditory feedback, and a test phase of 144 trials administered in a single block. Each trial began with the presentation of a fixation cross on an otherwise blank screen, after which dot arrays (yellow and blue dots) were simultaneously presented on the left and right sides of the screen for 500 ms. Participants were instructed to indicate which side contained more dots by pressing the 'f' key (left hand) or the 'j' key (right hand). Half of the trials featured the correct answer on the left side of the screen, the other half on the right. Numerosities ranged from 10 to 40 and six ratios were used: 1.11, 1.14, 1.2, 1.25, 1.5 and 2. The two tasks were completed sequentially, and the order was counterbalanced across participants. Accuracy scores were used for all analyses (see also Gilmore et al., 2014). Example items from each task are presented in Figure 1.

Panamath. One numerosity comparison task was created and administered using the Panamath Software (downloaded from the Panamath website; www.panamath.org). The version used here contained three trial types: *area-congruent* trials (i.e., trials in which the dot array that contained more dots had the larger cumulative surface area), *area-neutral* trials (i.e., trials in which both dot arrays had the same cumulative surface area), and *area-incongruent* trials (i.e., trials in which the dot array that contained more dots had the smaller cumulative surface area. Note that in this case, the cumulative perimeter between the dot arrays was equated).

Gebuis & Reynvoet (G&R). The other numerosity comparison task was an adapted version of the one used by Gebuis and Reynvoet (2011) and was administered using E-prime 3.0. The stimuli can be found on https://osf.io/rvh57. In this task, five visual cues were controlled for across trials: convex hull (i.e., the area subtended by each dot array), total surface area (i.e., the aggregate surface area of all dots in one array), dot item size (i.e., the average diameter of the dots presented in one array), total circumference (i.e., the aggregate circumference of all dots in one array), and density (i.e., surface area divided by convex hull). This task contained two trial

types: fully congruent and fully incongruent, meaning that all the visual cues were congruent during congruent trials and incongruent during incongruent trials.

Data Analysis Plan

Research questions, data collection, and a series of analyses for this investigation were preregistered on Open Science Framework on February 19, 2018 (osf.io/arwm5). Analyses were aimed at better understanding mixed results that have emerged from prior studies showing inconsistent relations between numerosity comparison and math achievement. Specifically, we address two research questions: (1) The extent to which two different numerosity comparison tasks are related to different math skills (arithmetic fluency, procedural calculation, word problem solving) and (2) whether different inhibitory control measures were related to performance on either assessment of numerosity comparison or mathematical skills.

We first presented a series of repeated measures analyses of variance (ANOVAs) to test for the effects of ratio and trial type in both numerosity comparison tasks and for the presence of interference effects in the Stroop tasks. Second, because the performance in both numerosity comparison tasks was affected by trial type, an Exploratory Factor Analysis (EFA) was conducted to uncover the underlying structure of the different trial types. This EFA was not preregistered but seemed a logical step in the data analysis given the strong effects of trial type. Finally, after computing descriptive statistics and bivariate relations for variables of interest, we ran two path models with all math assessments simultaneously regressed on each of the two number comparison tasks to address the first research question. To address our second research question, we ran a series of path models to test whether inhibitory control measures were independently related to performance on assessments of mathematical skill, and whether the relation between numerosity comparison and mathematical skills operates indirectly through inhibitory control. All path analyses were run in Mplus 8 (Muthén & Muthén, 2017).

Results

Within-task repeated measures ANOVAs.

Numerosity comparison. Two repeated-measures ANOVAs modeled six within-subject levels (one for each ratio). Additionally, each test modeled separate trial types (two levels for G&R: Congruent and incongruent; three levels for Panamath: area-congruent, area-neutral, and area-incongruent). Results of both ANOVAs are shown in Figure 2.

In the G&R numerosity comparison task, there was a main effect of trial type, $F(1, 121) = 371.76$, $p < .001$. Accuracies on congruent trials ($M = 0.88$, $SD = .09$) were higher than on incongruent trials ($M = 0.46$, $SD = .17$). There was also a main effect of ratio, $F(4.84,585.26) = 211.82$, $p < .001$, such that participants performed more accurately when presented with larger ratios. As expected, the interaction of ratio and congruency was also significant, $F(4.81, 582.28) = 67.68$, $p < .001$, such that participants improved more on incongruent trials with increasing ratios than they did on congruent trials. This is possibly due to particularly poor performance on low-ratio incongruent trials as compared to congruent trials which might have seen a ceiling effect; for trials that used a 1.11 ratio, accuracy on incongruent trials was near 30% (below chance), whereas for the same ratio in congruent trials, accuracy was above 80%.

For Panamath, there was also a main effect of trial type, $F(1.85, 224.34) = 21.09$, $p < .001$. Participants were more accurate on area-congruent trials ($M = 0.82$, $SD = .06$) than on area-neutral trials ($M = 0.80$, $SD = .07$) and area-incongruent trials ($M = 0.78$, $SD = .08$). Bonferroni post-hoc comparisons revealed that means of the three trial types were significantly different from one another (all p-values $< .01$). There was also a main effect of ratio, $F(4.43, 535.66) = 289.13$, $p < .001$, such that participants again performed worse on trials with smaller ratios. There was also a significant interaction of ratio and congruency, $F(8.19, 990.85) = 15.33$, $p < .001$. This again appeared to be driven by the fact that participants' accuracy for area-incongruent trials improved sharply as the ratio increased; for trials that used a 1.11 ratio, accuracy on area-incongruent trials was below 60%, whereas for area-congruent and area-neutral trials, accuracy was near 70%. Again, this might reflect a ceiling effect for easier trial types.

Stroop interference. Error rates were 15.35% and 3.17% respectively for the numerical Stroop and the animal Stroop task. Repeated measures ANOVAs were conducted for both Stroop

tasks to examine the expected presence of an interference effect. Results are presented in Figure 3. For the numerical Stroop task, there was a significant effect of congruency for RTs, $F(1,121) = 305.65$, $p < .001$. Participants were faster during congruent trials ($M = 487.32$ ms, $SD = 37.10$) compared to incongruent trials ($M = 529.46$ ms, $SD = 37.30$). The same was true for the animal Stroop task, $F(1,121) = 178.99$, $p < 0.001$, wherein participants were faster during congruent trials ($M = 351.55$ ms, $SD = 39.69$) than incongruent trials ($M = 380.49$ ms, $SD = 47.33$). There was also a significant difference in accuracy between the two trial types in both number and animal Stroop tasks, $F(1,121) = 61.00$, $p < .001$; $F(1,121) = 233.15$, $p < .001$, respectively. In both tasks, participants were more accurate in congruent trials than in incongruent trials. However, only an interference score for RTs was computed and used for all subsequent analyses due to the likelihood of ceiling effects on the congruent trials, as the means for both Stroop tasks were above 90%.

Exploratory Factor Analysis for Numerosity Comparison tasks.

Because both analyses of the numerosity comparison tasks revealed an effect of trial type, an exploratory factor analysis was conducted to determine the underlying factor structure of the different trials. Trial accuracy was aggregated to the ratio bin level for each trial type in the two tasks. That is, a total of 30 variables were included, 12 representing accuracy on the G&R incongruent or congruent trials in each of the six ratio bins and 18 representing accuracy on the three Panamath trial types in each of the six ratio bins. These variables were submitted to an EFA with ROTATION = OBLIMIN. A three-factor model fit the data significantly better than did a two-factor model, $\chi^2(28) = 50.399$, $p = .0058$, and a four-factor model did not fit the data better than did a three-factor model, $\chi^2(27) = 30.441$, $p = .2947$, suggesting that a three-factor model was most appropriate. However, this model still did not fit the data well according to conventional fit statistics, $\chi^2(348) = 513.587$, $p < .0001$, RMSEA = .062 90% CI[.051, .074], CFI = 0.836, SRMR = .059, suggesting that the variance in these 30 variables could not be well reduced to a small set of factors, possibly due to extraneous error in these composites created from a relatively low number of trials within each bin (8 trials per bin for Panamath, 12 trials per bin for G&R). As such, these latent factors were not used in structural models and instead the pattern of item loadings was examined more descriptively. Item loadings onto the three factors are shown in Table 1. All variables representing congruent trials from the G&R assessment

loaded onto one factor, all variables representing incongruent trials loaded onto another, and most variables representing trials from Panamath loaded onto a third. Thus, for subsequent analyses we use three separate indices of performance on the numerosity comparison tasks: overall Panamath accuracy, G&R congruent accuracy, and G&R incongruent accuracy.

Path Analyses.

Basic descriptive statistics of the variables of interest for the path analyses are presented in Table 2. Bivariate correlations among all relevant variables are presented in Table 3.

Numerosity comparison tasks and math tests. To test whether numerosity comparison skills are related to mathematical skills, two separate path models were conducted wherein the three math skills (arithmetic fluency, procedural calculation, and applied word problems) were simultaneously regressed on each of the three measures of numerosity comparison skills by task (i.e., Panamath in one model; G&R in the other). Covariances among all outcome variables were modeled such that all models were fully identified so as to assess the unique contribution of numerosity comparison skills to each outcome measure net of any similarity between different measures of mathematical skills. Results of both path models are shown in Figure 4. Accuracy on the Panamath task was related to relative success on both procedural calculation, $\beta = .17$, $p = .039$, and word problems, $\beta = .19$, $p = .026$. In contrast, there was no relation between accuracy on either incongruent or congruent trials on the G&R numerosity comparison task and any math skills. A sensitivity test in which all math skills were regressed on both G&R congruent and G&R incongruent accuracy in separate models demonstrated that neither alone was related to math skills either.

To test whether associations between numerosity comparison and various mathematical skills differed depending on the numerosity comparison task utilized (e.g., whether accuracy on the Panamath task was more strongly related to arithmetic fluency than was accuracy on the G&R incongruent trials), we then conducted a multi-group path model with each numerosity measure (i.e., Panamath, G&R incongruent, G&R congruent) representing a different group. Parameters were tested for equality across the three groups; a Wald test confirmed parameters were not significantly different across numerosity measures in predicting any of the three math outcomes, arithmetic fluency: $\chi^2(2) = 1.87$, $p = .392$; procedural calculation: $\chi^2(2) = 5.71$, $p = .058$; applied

word problems: $\chi^2(2) = 2.85$, $p = .240$. Due to the fact that the analysis of parameter equality on procedural calculation neared conventional levels of statistical significance, further post-hoc tests on group differences were performed. A comparison of G&R congruent and incongruent trials revealed pathways predicting procedural calculation from the G&R conditions were not significantly different from one another, $\chi^2(1) = 1.68$, $p = .195$. However, pathways predicting procedural calculation from the Panamath task were marginally different from the G&R congruent trials, $\chi^2(1) = 2.74$, $p = .098$, and the G&R incongruent trials, $\chi^2(1) = 3.22$, $p = .073$, such that the magnitude of these associations was larger for the Panamath task than for the two G&R trial types.

Lastly, to test whether there was a unique contribution of accuracy on the Panamath numerosity comparison task net of any covariance shared with either congruent or incongruent trials from the G&R numerosity comparison task, a single path model was conducted in which the three math assessments were simultaneously regressed on the three numerosity comparison measures. Panamath continued to predict procedural calculation ($\beta = .17$, $p = .039$) and marginally predict applied word problems ($\beta = .17$, $p = .068$), whereas neither congruent nor incongruent trials from the G&R task were related to math outcomes, suggesting Panamath measures a construct that is uniquely related to certain mathematical skills.

Inhibition, numerosity comparison, and math tests. Because of the strength of the correlation ($r = .60$) between the two Go/No-Go tasks, an aggregate was created to represent the total number of commission errors. A path model was run with the three math skills simultaneously regressed on the three inhibition scores (i.e., the Go/No-Go commission error composite, Number Stroop interference, and Animal Stroop interference), results of which are shown in Figure 5a. Only the interference score from the Animal Stroop task—but neither the Number Stroop nor total commission errors on Go/No-Go tasks—was related to performance on any math measure. At that, performance on the Animal Stroop task was only related to performance on the measure of arithmetic fluency, $\beta = .20$, $p = .021$.

Another path model (Figure 5b) was run to test whether Panamath continued to relate to math outcomes over and above inhibitory skills that were related to math outcomes. That is arithmetic fluency, procedural calculation, and applied word problems were simultaneously regressed on Panamath and the interference score on the Animal Stroop task. Similar to the

model without inhibitory skills, Panamath continued to relate to applied word problems, β = .19, p = .026, and procedural calculation, β = .17, p = .037. Inhibition continued to predict arithmetic fluency, β = .18, p = .050. There were no indirect associations of numerosity comparison skills with math skills via inhibitory control skills, which is not surprising given that Panamath and inhibitory control each related to different math skills.

Finally, a last path model was run with all parameters simultaneously estimated (i.e., both inhibition scores and all three ANS scores predicting all three math outcome measures) to test specificity and robustness of effects. With this more saturated model, we find only one significant association between RT on the Animal Stroop task and the test of arithmetic fluency, β = .20, p = .022, consistent with other models. Panamath remained marginally related to applied word problems, β = .15, p = .084, and procedural calculation, β = .15, p = .077. Neither G&R nor any of the other inhibition tasks was related to any outcome measure, ps > .163. This general lack of other findings potentially reflects an issue of low power or limited variance when all parameters are estimated simultaneously.

Discussion

The relation between numerosity comparison and math achievement has been the topic of intense debate. The discussion has been complicated by the fact that studies have used different math assessments and numerosity comparison tasks that used different stimulus generation algorithms. These tasks may pose greater or lesser demands on inhibition, which has been measured in past work in a variety of ways as well. Here, we examine the influence of the stimulus generation algorithm used in different numerosity comparison tasks, the type of math assessment, and the role that inhibition plays on the relation between numerosity comparison and math. More specifically, we compared two of the most commonly utilized algorithms (Panamath, Halberda et al., 2008; G&R, Gebuis & Reynvoet, 2011) and their relation with arithmetic fluency, procedural calculation, and applied word problem solving skills and whether different forms of inhibition (measured via Stroop and Go/No-Go tasks) influence the relation between numerosity comparison tasks and math skills. Before going into the main results of our study, we want to repeat that, due to the removal of some outliers, our final sample (n = 122) was a bit smaller than the number required based on our a-priori power analyses (n = 133), which may

have had a minor impact on some of the results. Nevertheless, our results suggest that the two numerosity tasks measure different cognitive processes which relate differently to math and to inhibition. We discuss each of these results in greater detail below.

First, correlations between performance on the two numerosity comparison tasks are dependent on trial type. There is a weak but significant correlation between Panamath and the G&R incongruent trials (r = .25), but no correlation between Panamath and the G&R congruent trials (r = .01). If both trials types of the G&R tasks are aggregated, the correlation with the Panamath version is r = .33, an effect size that is very similar to those observed in previous studies by Clayton et al. (2015) and DeWind and Brannon (2016). The strongest correlation that was observed between trial types, although negative, was observed between G&R congruent and G&R incongruent trials (r = -.60), indicating that there is a trade-off between both trials of the same task: better performance on congruent trials is associated with worse performance on incongruent trials. Interestingly, our exploratory factor analysis confirmed that the cognitive processes indexed by the Panamath task that consisted of area-congruent, area-neutral, and area-incongruent trials were more closely related to each other than to those indexed by the G&R incongruent and congruent trials. These findings suggest that the stimulus generation algorithms have a major impact on the performance in numerosity comparison tasks; however, further study is needed with larger pools of items to better understand the underlying structure of tasks.

To arrive at a possible explanation for these observations, we first want to reiterate some details of both stimulus generation algorithms. As mentioned before, we used three different trial types in the Panamath numerosity comparison task: trials wherein the cumulative surface area of all dots in an array and number was positively correlated, trials wherein they were equated, and trials wherein they were negatively correlated and cumulative perimeter was in turn equated. While most studies using Panamath only use two trial types (i.e., area-congruent and area-neutral trials), we also included area-incongruent trials because previous work has found that some participants may use total perimeter as a non-numerical cue to solve non-symbolic number comparison tasks (DeWind, Adams, Platt & Brannon, 2015). The G&R version that was used in this study contains two different trials types: fully congruent and fully incongruent, in which all non-numerical cues (convex hull, total surface area, dot item size, total circumference, and density) were respectively congruent or incongruent with number (see Figure 1 for examples of these trial types). Many studies have confirmed that participants' numerosity comparisons are

biased by non-numerical information present in the dot arrays, as evidenced by the congruency effects (e.g., Clayton et al., 2015; Defever et al., 2013; Fuhs & McNeil, 2013; Leibovich, & Henik, 2014; Norris, Clayton, Gilmore, Inglis, & Castronovo, 2019; Reynvoet et al., 2018). The present results are in line with this pattern of findings and reveal congruency effects in both numerosity comparison tasks. Remarkably, the congruency effect is much larger in the G&R version of the task than in the Panamath version. More specifically, in the G&R numerosity comparison task, participants' accuracy is below chance on incongruent trials with difficult ratios, whereas average performance on congruent trials does not fall below .80 for any ratio. This is an indication that in the G&R task, participants' decisions are strongly influenced by non-numerical information.

A possible explanation as to why participants rely more on non-numerical cues in some cases than others is offered by the signal clarity account (Cantrell, Boyer, Cordes, & Smith, 2015; see also Leibovich et al., 2017). This account posits that dimensions which contain a lot of variance (i.e., a salient dimension) are more easily extracted and used in decisions. DeWind and Brannon (2016; Table 1) systematically analyzed all non-numerical cues of both stimulus generation algorithms and found that, although non-numerical cues correlate less with number in the G&R algorithm than in the Panamath algorithm (i.e., convex hull/spacing is not controlled in this task, Dewind & Brannon, 2016), the variance of the non-numerical cues in the G&R algorithm is much larger than in Panamath, making these non-numerical cues more salient and as a consequence, more influential in numerical decisions. In other words, non-numerical dimensions may have had a greater impact on decisions in the G&R numerosity comparison task than in the Panamath version, at the cost of the impact of the number dimension.

Our second main finding was that performance on the two numerosity comparison tasks was differentially related to our math assessments. Performance on the G&R numerosity comparison task was unrelated to any aspect of math, whereas performance on the Panamath numerosity comparison task related to measures of procedural calculations and applied word problem solving but not arithmetic fluency, supporting prior findings (e.g., Schneider et al., 2017). One possible explanation for why Panamath and not the G&R task is related to math outcomes is that because of the larger diversity in the trials types (area-congruent, area-neutral, area-incongruent) stimuli in the Panamath task, number is a more prominent feature in the Panamath task than in the G&R task. These numerical representations may in turn be of

particular importance for procedural calculation and word problem solving because these aspects of math rely much more on numerical representations than arithmetic fluency which relies most heavily on memory retrieval. An alternative explanation for why Panamath may be related to math outcomes is that both Panamath and the math tests require some degree of cognitive flexibility. The Panamath task can indeed be accurately performed by flexibly switching between different decision strategies (e.g., Roquet & Lemaire, 2019). For instance, in area-congruent trials, a correct decision can be derived on the basis of area. In area-neutral trials, the array with the smaller dots is the more numerous one and as a consequence, the correct response can be derived on the basis of individual dot size. Finally, in area-incongruent trials, participants may be able to use the same strategy as on area-congruent trials but reverse their response. Thus, participants who are more flexible at switching between these strategies may perform better on the task overall. Also for most math tests, cognitive flexibility is a requirement as participants need to be able to switch between different strategies or algorithms to solve the task successfully (e.g., Hodzik & Lemaire, 2011). However, we think that it is unlikely that a shared need for cognitive flexibility may explain the link between Panamath performance and math ability because the type of cognitive flexibility involved is different in both tasks (i.e., flexibly changing between perceptual inputs vs flexibly adapting strategies/algorithms in math). A recent meta-analysis on far-transfer effects of training components of children's executive functions skills (including cognitive flexibility) found no convincing evidence for such far-transfer effects (Kasai, Futo, Demetrovics, & Takacs, 2019). Thus, it seems unlikely that variations in the use of cognitive flexibility on the Panamath task would relate to cognitive flexibility on the math assessment. However, to completely rule out this alternative possibility future studies should include a task measuring cognitive flexibility as an additional control task. If cognitive flexibility indeed explains the relation between Panamath and math outcomes, the relation should no longer be significant when cognitive flexibility is controlled for.

Our findings may explain the contradicting findings from previous studies that examined the relation between numerosity comparison and mathematics achievement (e.g. Halberda et al., 2012; Libertus et al., 2012; Libertus et al., 2011; Sasanguie et al., 2014; Sasanguie et al., 2013) and the large heterogeneity observed in meta-analyses on this relation (Chen & Li, 2014; Schneider et al., 2017). Apparently, the relation between numerosity comparison and math achievement is influenced by the way in which stimuli for the numerosity comparison task are

constructed *and* by the way math achievement is measured. These findings partially replicate previous findings by Braham and Libertus (2018), who showed that performance on a numerosity comparison task using stimuli created with the Panamath algorithm correlated with applied word problem solving. However, Braham and Libertus found a significant association between numerosity comparison and math fluency but not calculation, whereas in these analyses we observed that numerosity comparison was related to calculation but not math fluency. These differences may in part be due to slight variations between the math fluency and calculation measures used. For example, the math fluency measure used by Braham and Libertus mixes arithmetic operations while the majority of problems on the math fluency measure used in the present study are blocked by arithmetic operation. Another potential explanation is that in the current study, the fluency measure was based on relatively more fact retrieval problems than in the study by Braham and Libertus because the first items in the present arithmetic fluency task are predominantly fact retrieval problems. In addition, the calculation measure used by Braham and Libertus includes more advanced concepts such as fractions, algebra, geometry, trigonometry, and factorials. Alternatively, Braham and Libertus showed that math anxiety moderated the relation between numerosity comparison and applied word problem solving performance, suggesting that other characteristics of the participants and the sample as a whole may explain contradicting findings in the literature. In that respect, it is important to emphasize again that our sample was rather homogeneous and consisted of university students and that these results might be slightly different in more heterogeneous populations.

Our third main finding was that the two numerosity comparison tasks related differently to our four inhibition measures. Performance on the Panamath task was significantly correlated with the commission errors made in both the Number and Animal Go/No-Go task and with the interference effect observed in the Number Stroop task. In contrast, the G&R numerosity comparison task was not correlated with any of the inhibition measures. One possible explanation is that participants were not actively trying to inhibit non-numerical information in the G&R task, while the Panamath task may have triggered a greater reliance on inhibitory control to extract the numerical information needed to complete the task. Importantly, inhibition was not responsible for the correlation between the Panamath numerosity comparison task and math achievement as the associations between these two variables remained significant when controlling for inhibition. This result is in line with the findings of Keller and Libertus (2015) in young children but

contradicts findings of Gilmore et al. (2013) and Fuhs and McNeil (2013). Finally, even when controlling for performance on the G&R numerosity comparison task, individual differences in accuracy on the Panamath numerosity comparison task continue to predict math achievement, specifically procedural calculation.

In conclusion, this study replicates and extends much of the prior research regarding the relations between non-symbolic number sense and symbolic number understanding and provides some further clarity to seemingly inconsistent findings. This study is the first to systematically examine the impact of variations in math assessment, stimulus generation algorithms to assess numerosity comparison skills, and inhibition for the link between numerosity comparison and math achievement. Our results suggest that variations in the stimulus generation protocols for numerosity comparison tasks result in different reliance on non-numerical cues and non-symbolic number representations. Possibly, non-symbolic number is processed more prominently in the Panamath task because of the larger variance in trial types that are presented during its administration, while the large variation in non-numerical cues in the G&R algorithm results in decisions heavily influenced by non-numerical cues. However, alternative explanations underlying the relation between Panamath and math tests such as cognitive flexibility need to be examined further. This explanation fits with our findings that performance on the Panamath task but not the G&R task was correlated with math achievement and that these results held even when controlling for inhibition. In sum, our findings with adults explain prior mixed findings regarding the link between non-symbolic number sense and math and highlight the need to carefully consider the stimulus generation protocols used for numerosity comparison tasks and the choice of math measures. Future studies should explore how these factors impact the link between non-symbolic number sense and math through development.

Funding

This research was supported by a grant from the Fund for Scientific Research- Flanders awarded to Bert Reynvoet and Delphine Sasanguie.

References

Agrillo, C., Petrazzini, M. E. M., & Bisazza, A. (2016). Number versus continuous quantities in lower vertebrates. In Henik, A. (Ed.). *Continuous Issues in Numerical Cognition* (pp. 149-174). Academic Press.

Braham, E. J., & Libertus, M. E. (2018). When approximate number acuity predicts math performance: The moderating role of math anxiety. *PloS One, 13*(5).

Brankaer, C., Ghesquière, P., & De Smedt, B. (2014). Children's mapping between non-symbolic and symbolic numerical magnitudes and its association with timed and untimed tests of mathematics achievement. *PloS One*, *9*(4).

Cantrell, L., Boyer, T. W., Cordes, S., & Smith, L. B. (2015). Signal clarity: An account of the variability in infant quantity discrimination tasks. *Developmental Science, 18*(6), 877-893.

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.

Clayton, S., Gilmore, C., & Inglis, M. (2015). Dot comparison stimuli are not all alike: The effect of different visual controls on ANS measurement. *Acta Psychologica*, *161*, 177-184.

Davis, J. L., & Matthews, R. N. (2010). NEPSY-II Review: Korkman, M., Kirk, U., & Kemp, S.(2007). NEPSY—Second Edition (NEPSY-II). San Antonio, TX: Harcourt Assessment. *Journal of Psychoeducational Assessment*, *28*(2), 175-182.

Defever, E., Reynvoet, B., & Gebuis, T. (2013). Task-and age-dependent effects of visual stimulus properties on children's explicit numerosity judgments. *Journal of Experimental Child Psychology*, *116*(2), 216-233.

Dehaene, S. (2011). *The number sense: How the mind creates mathematics*. Oxford University Press.

Desoete, A., & Roeyers, H. (2006). Cognitieve deelvaardigheden rekenen. *Herentals: Vlaamse Vereniging voor Logopedisten*.

DeWind, N. K., Adams, G. K., Platt, M. L., & Brannon, E. M. (2015). Modeling the approximate number system to quantify the contribution of visual stimulus features. *Cognition*, *142*, 247-265.

DeWind, N. K., & Brannon, E. M. (2016). Significant inter-test reliability across approximate number system assessments. *Frontiers in Psychology*, *7*, 310.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175-191.

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.

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.

Fuhs, M. W., McNeil, N. M., Kelley, K., O'Rear, C., & Villano, M. (2016). The role of non-numerical stimulus features in approximate number system training in preschoolers from low-income homes. *Journal of Cognition and Development*, *17*(5), 737-764.

Gebuis, T., & Reynvoet, B. (2011). Generating nonsymbolic number stimuli. *Behavior Research Methods*, *43*(4), 981-986.

Gebuis, T., & Reynvoet, B. (2012a). The role of visual information in numerosity estimation. *PloS One*, *7*(5).

Gilmore, C., Attridge, N., Clayton, S., Cragg, L., Johnson, S., Marlow, N., Simms, V. & Inglis, M. (2013). Individual differences in inhibitory control, not non-verbal number acuity, correlate with mathematics achievement. *PloS One*, *8*(6).

Gilmore, C., Attridge, N., De Smedt, B., & Inglis, M. (2014). Measuring the approximate number system in children: exploring the relationships among different tasks. *Learning and Individual Differences*, *29*, 50-58.

Gilmore, C., Cragg, L., Hogan, G., & Inglis, M. (2016). Congruency effects in dot comparison tasks: Convex hull is more important than dot area. *Journal of Cognitive Psychology*, *28*(8), 923-931.

Halberda, J., & Feigenson, L. (2008). Developmental change in the acuity of the" Number Sense": The Approximate Number System in 3-, 4-, 5-, and 6-year-olds and adults. *Developmental Psychology*, *44*(5), 1457.

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

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*, *109*(28), 11116-11120.

Hodzik, S., & Lemaire, P. (2011). Inhibition and shifting capacities mediate adults' age-related differences in strategy selection and repertoire. *Acta Psychologica*, *3*, 335-344.

Jonkman, L. M. (2006). The development of preparation, conflict monitoring and inhibition from early childhood to young adulthood; a Go/Nogo ERP study. *Brain research*, *1097*(1), 181-193.

Kasai, R., Futo, J., Demetrovics, Z., & Takacs, Z.K. (2019). A meta-analysis of the experimental evidence on the near- and far-transfer effects among children's executive function skills. *Psychonomic Bulletin, 145*, 165-188.

Keller, L., & Libertus, M. (2015). Inhibitory control may not explain the link between approximation and math abilities in kindergarteners from middle class families. *Frontiers in Psychology*, *6*, 685.

Leibovich, T., & Henik, A. (2014). Comparing performance in discrete and continuous comparison tasks. *The Quarterly Journal of Experimental Psychology*, *67*(5), 899-917.

Libertus, M. E., & Brannon, E. M. (2010). Stable individual differences in number discrimination in infancy. *Developmental Science*, *13*(6), 900-906.

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.

Libertus, M. E., Odic, D., Feigenson, L., & Halberda, J. (2016). The precision of mapping between number words and the approximate number system predicts children's formal math abilities. *Journal of Experimental Child Psychology*, *150*, 207-226.

Libertus, M. E., Odic, D., Feigenson, L., & Halberda, J. (2020). Effects of Visual Training of Approximate Number Sense on Auditory Number Sense and School Math Ability. *Frontiers in Psychology*, *11*, 2085.

Libertus, M. E., Odic, D., & Halberda, J. (2012). Intuitive sense of number correlates with math scores on college-entrance examination. *Acta Psychologica*, *141*(3), 373-379.

Mazzocco, M. M., Feigenson, L., & Halberda, J. (2011). Impaired acuity of the approximate number system underlies mathematical learning disability (dyscalculia). *Child Development*, *82*(4), 1224-1237.

Muthén, L. K., & Muthén, B. O. (2017). 1998–2017. Mplus user's guide. *Los Angeles, CA: Muthén & Muthén*.

Norris, J. E., & Castronovo, J. (2016). Dot display affects approximate number system acuity and relationships with mathematical achievement and inhibitory control. *PloS One*, *11*(5).

Norris, J. E., Clayton, S., Gilmore, C., Inglis, M., & Castronovo, J. (2019). The measurement of approximate number system acuity across the lifespan is compromised by congruency effects. *Quarterly Journal of Experimental Psychology*, *72*(5), 1037-1046.

Nys, J., & Content, A. (2012). Judgement of discrete and continuous quantity in adults: Number counts!. *The Quarterly Journal of Experimental Psychology*, *65*(4), 675-690.

Odic, D., Hock, H., & Halberda, J. (2014). Hysteresis affects approximate number discrimination in young children. *Journal of Experimental Psychology: General*, *143*(1), 255.

Odic, D., Libertus, M. E., Feigenson, L., & Halberda, J. (2013). Developmental change in the acuity of approximate number and area representations. *Developmental Psychology*, *49*(6), 1103.

Park, J., Bermudez, V., Roberts, R. C., & Brannon, E. M. (2016). Non-symbolic approximate arithmetic training improves math performance in preschoolers. *Journal of Experimental Child Psychology*, *152*, 278-293.

Park, J., & Brannon, E. M. (2013). Training the approximate number system improves math proficiency. *Psychological Science*, *24*(10), 2013-2019.

Park, J., & Brannon, E. M. (2014). Improving arithmetic performance with number sense training: An investigation of underlying mechanism. *Cognition*, *133*(1), 188-200.

Piazza, M., De Feo, V., Panzeri, S., & Dehaene, S. (2018). Learning to focus on number. *Cognition*, *181*, 35-45.

Piazza, M., Facoetti, A., Trussardi, A. N., Berteletti, I., Conte, S., Lucangeli, D., Dehaene, S. & Zorzi, M. (2010). Developmental trajectory of number acuity reveals a severe impairment in developmental dyscalculia. *Cognition*, *116*(1), 33-41.

Piazza, M., Pica, P., Izard, V., Spelke, E. S., & Dehaene, S. (2013). Education enhances the acuity of the nonverbal approximate number system. *Psychological Science*, *24*(6), 1037-1043.

Reynvoet, B., Vos, H., & Henik, A. (2019). Comparative judgment of familiar objects is modulated by their size. *Experimental Psychology*, *65*, 353-359.

Roquet, A., & Lemaire, P. (2019). Strategy variability in numerosity comparison task: a study in young and older adults. *Open Psychology*, *1*, 152-167.

Rousselle, L., & Noël, M. P. (2008). The development of automatic numerosity processing in preschoolers: Evidence for numerosity-perceptual interference. *Developmental Psychology*, *44*(2), 544.

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

Sasanguie, D., De Smedt, B., & Reynvoet, B. (2017). Evidence for distinct magnitude systems for symbolic and non-symbolic number. *Psychological Research*, *81*(1), 231-242.

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.

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.

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*, *20*(3), e12372.

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.

Smets, K., Gebuis, T., Defever, E., & Reynvoet, B. (2014). Concurrent validity of approximate number sense tasks in adults and children. *Acta Psychologica*, *150*, 120-128.

Smets, K., Moors, P., & Reynvoet, B. (2016). Effects of presentation type and visual control in numerosity discrimination: implications for number processing. *Frontiers in Psychology*, *7*, 66.

Starr, A., Libertus, M. E., & Brannon, E. M. (2013). Number sense in infancy predicts mathematical abilities in childhood. *Proceedings of the National Academy of Sciences*, *110*(45), 18116-18120.

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

Viswanathan, P., & Nieder, A. (2015). Differential impact of behavioral relevance on quantity coding in primate frontal and parietal neurons. *Current Biology*, *25*(10), 1259-1269.

de Vos, T. (1992). *Tempo-Test-Rekenen: test voor het vaststellen van het rekenvaardigheidsniveau der elementaire bewerkingen (automatisering) voor het basis-en voortgezet onderwijs: handleiding*. Berkhout.

Xu, F., Spelke, E. S., & Goddard, S. (2005). Number sense in human infants. *Developmental Science*, *8*(1), 88-101.

Table 1. Results of exploratory factor analysis

	% Correct	SD	Factor 1	Factor 2	Factor 3
Panamath Area Correlated 1.11	70.33	13.87	0.087	-0.178	-0.048
Panamath Area Correlated 1.14	74.66	13.34	0.328	-0.150	0.055
Panamath Area Correlated 1.20	83.62	12.95	0.358*	0.013	-0.206
Panamath Area Correlated 1.25	77.99	12.55	0.361	0.245	0.269
Panamath Area Correlated 1.50	91.07	11.04	0.438*	-0.064	0.204
Panamath Area Correlated 2.00	98.41	5.20	0.422	-0.019	0.079
Panamath Area Equal 1.11	68.77	17.71	0.099	0.070	0.128
Panamath Area Equal 1.14	71.43	16.12	0.249	-0.038	-0.129
Panamath Area Equal 1.20	71.50	15.75	0.299	0.199	0.129
Panamath Area Equal 1.25	84.55	12.77	0.367	-0.027	0.169
Panamath Area Equal 1.50	89.90	10.64	0.451*	0.102	0.135
Panamath Area Equal 2.00	96.52	6.58	0.416*	0.035	0.090
Panamath Area Anticorrelated 1.11	56.02	15.90	0.084	0.043	0.166
Panamath Area Anticorrelated 1.14	70.20	16.89	0.444*	-0.029	-0.103
Panamath Area Anticorrelated 1.20	76.37	14.46	0.356*	0.106	-0.026
Panamath Area Anticorrelated 1.25	84.87	12.08	0.329	-0.018	-0.019
Panamath Area Anticorrelated 1.50	86.27	14.85	0.533*	0.002	-0.235*
Panamath Area Anticorrelated 2.00	93.14	12.14	0.299	0.155	-0.179
G&R Congruent 1.11	82.98	14.13	-0.072	-0.263	0.550*
G&R Congruent 1.14	83.83	13.94	0.012	-0.305	0.566*
G&R Congruent 1.20	84.37	16.14	0.024	-0.229	0.533*
G&R Congruent 1.25	87.32	12.05	-0.036	0.018	0.762*
G&R Congruent 1.50	91.06	10.19	0.034	0.090	0.685*
G&R Congruent 2.00	95.52	7.22	0.083	0.069	0.624*
G&R Incongruent 1.11	31.68	20.02	0.038	0.640*	-0.232
G&R Incongruent 1.14	31.48	18.32	-0.181	0.840*	0.044
G&R Incongruent 1.20	36.78	19.88	0.002	0.512*	-0.360*
G&R Incongruent 1.25	45.26	21.66	0.063	0.709*	-0.054
G&R Incongruent 1.50	58.57	22.80	0.048	0.829*	-0.030
G&R Incongruent 2.00	73.86	22.11	0.174	0.777*	0.080

*Note: * factor loading significant at p < .05; G&R—Gebuis & Reynvoet*

Table 2. Descriptive statistics for variables included in path analyses.

	Mean	**SD**	**Minimum**	**Maximum**
Age	20.23	2.05	17.89	29.40
G&R Congruent Accuracy	0.88	0.09	0.63	1.00
G&R Incongruent Accuracy	0.46	0.17	0.10	0.85
Panamath Total Accuracy	0.80	0.05	0.67	0.91
Number Go/No-Go Commission Errors	6.87	3.60	1.00	16.00
Animal Go/No-Go Commission Errors	5.05	3.43	0.00	16.00
Number Stroop Interference RT Score	42.14	26.62	-14.00	123.00
Animal Stroop Interference RT Score	28.93	23.89	-13.00	108.50
Arithmetic Fluency	104.86	17.45	66	151
Procedural Calculation Sum Score	6.69	1.91	2	10
Applied Mathematics Sum Score	5.89	1.91	1	10

Table 3. Bivariate correlations among variables included in path analyses

	1	2	3	4	5	6	7	8	9	10	11
1. Age	—										
2. G&R Congruent Accuracy	.06	—									
3. G&R Incongruent Accuracy	-.04	-.60**	—								
4. Panamath Total Accuracy	.05	.01	.25**	—							
5. Number Go/No-Go Commission Errors	-.04	-.11	-.09	-.21*	—						
6. Animal Go/No-Go Commission Errors	-.04	-.02	-.18	-.24**	.60**	—					
7. Number Stroop Interference RT Score	.12	-.07	-.00	.19*	-.15	-.11	—				
8. Animal Stroop Interference RT Score	-.01	.07	-.16	-.07	-.03	.03	-.15	—			
9. Arithmetic Fluency	.11	-.13	.02	-.05	-.09	.00	.07	.19*	—		
10. Procedural Calculation Sum Score	.11	-.09	.10	.17	-.07	-.09	.08	.01	.46**	—	
11. Applied Mathematics Sum Score	.14	.01	.08	.19*	-.07	-.01	.11	-.01	.25**	.24**	—

Figure 1. Example stimuli of both algorithms (top: Panamath; below: G&R)

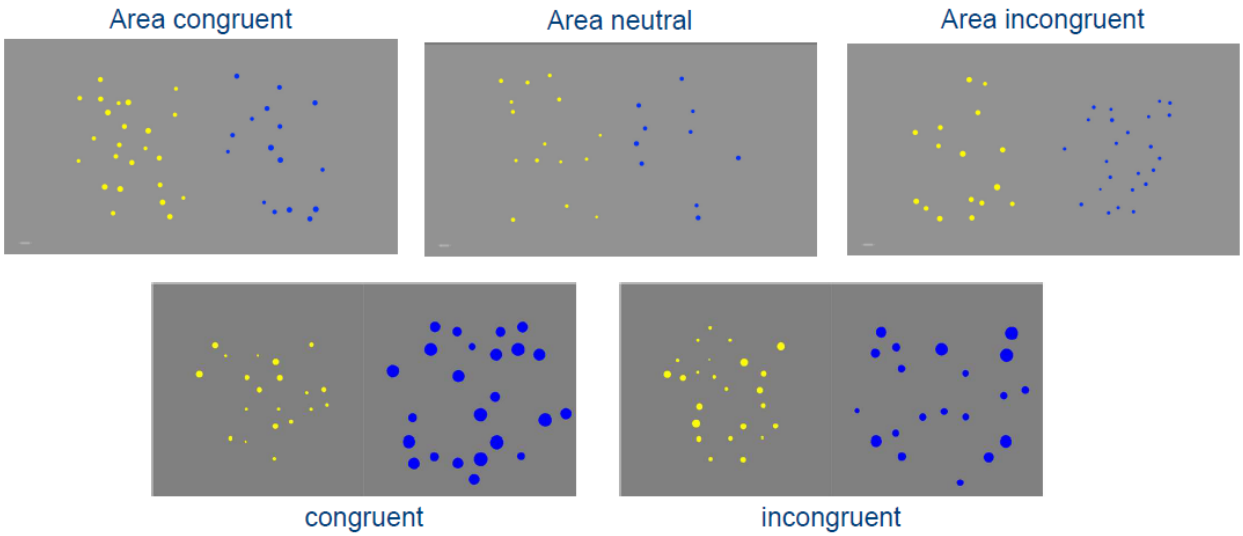
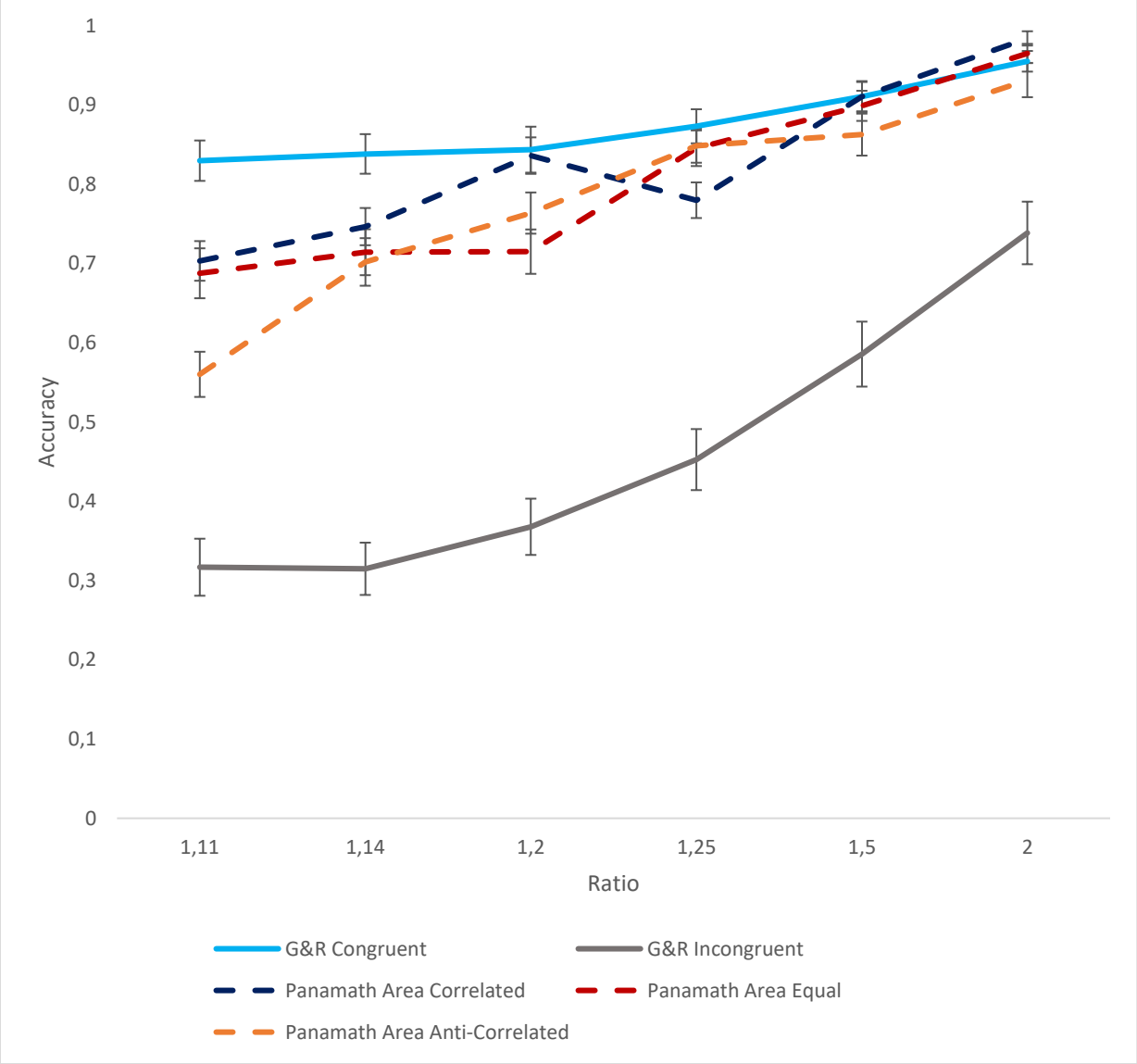
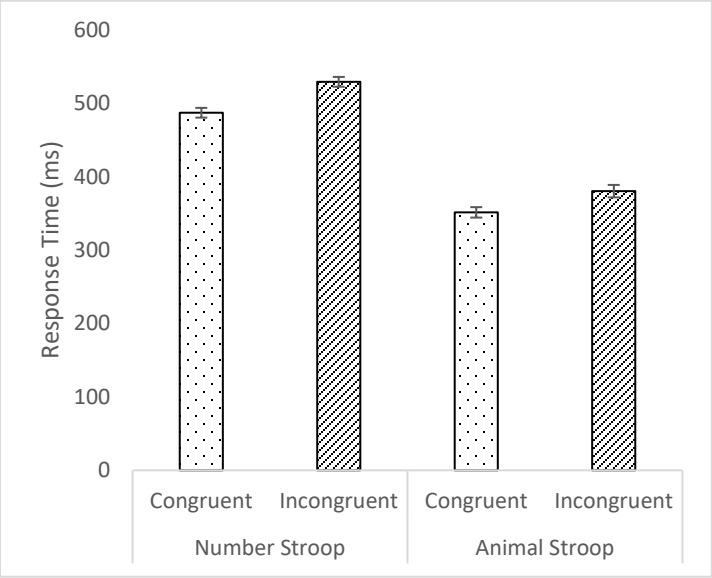


Figure 2. Mean accuracy on the two numerosity comparison tasks by ratio and trial type



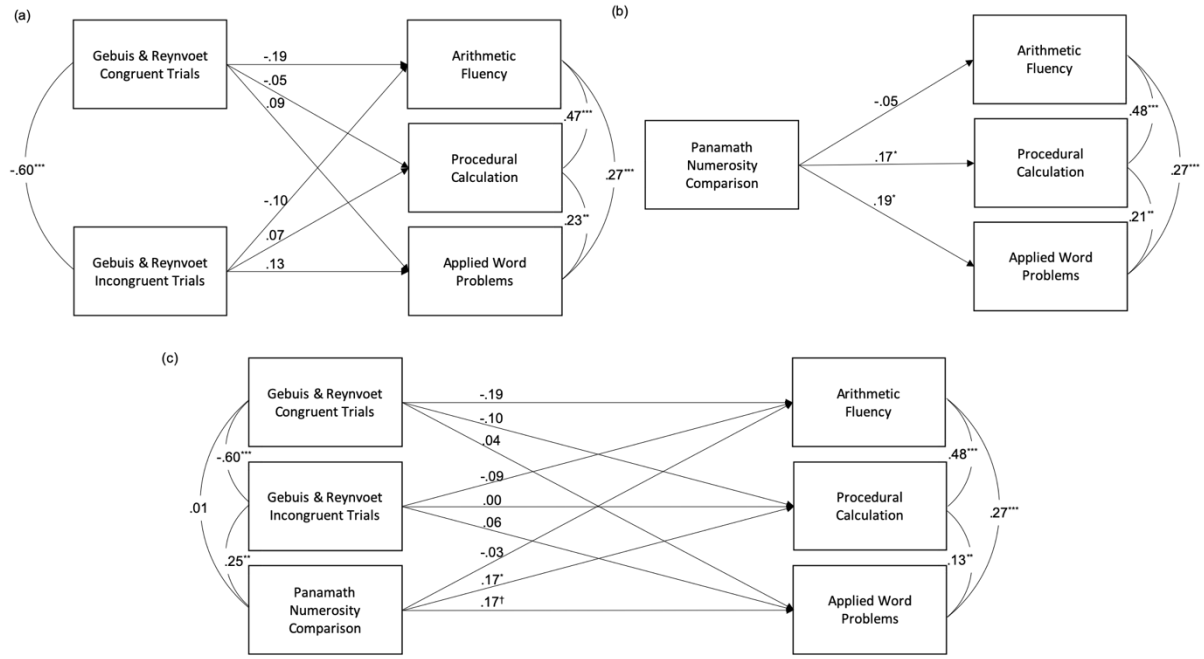
Note: Error bars represent 95% CIs.

Figure 3. Mean response time by task and trial type for the two Stroop tasks.



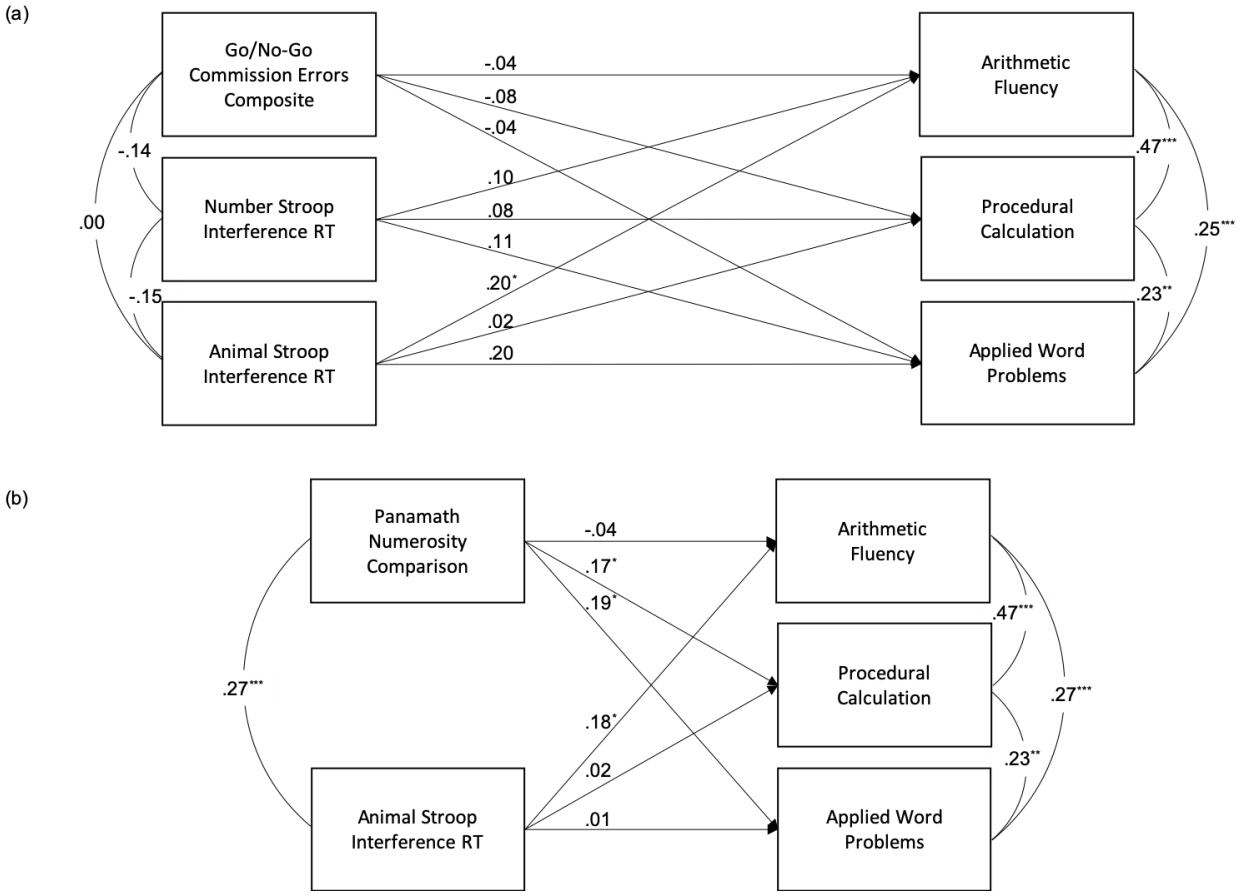
Note: Error bars represent 95% CIs.

Figure 4. Path analyses predicting math skills from numerosity comparison tasks



Note: † $p < .07$, * $p < .05$, ** $p < .01$, *** $p < .001$. All values represent standardized coefficients (a) Path model with math skills simultaneously regressed on Gebuis & Reynvoet numerosity comparison task; (b) Path model with math skills simultaneously regressed on Panamath numerosity comparison task; (c) Path model with math skills simultaneously regressed on both Gebuis & Reynvoet and Panamath numerosity comparison tasks.

Figure 5. Path analyses predicting math skills from inhibitory tasks



Note: * $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$. (a) Path model with math skills simultaneously regressed on all inhibition measures; (b) Path model with math skills simultaneously regressed on significant predictors of math from prior models.