Sharpening, focusing, and developing: a study of change in nonsymbolic number comparison skills and math achievement in 1st grade

Eric D. Wilkey[1], Lina Shanley[2], Fred Sabb[2], Daniel Ansari[1], Jason C. Cohen[2], Virany Men[2], Nicole A. Heller[2], and Ben Clarke[2]

[1] Brain & Mind Institute, Western University, London, Ontario, Canada

[2] University of Oregon, Eugene, Oregon, USA

Corresponding Author:

Ben Clarke
University of Oregon
1600 Milrace Dr.
Suite 207
Eugene, OR 97403, USA
clarkeb@uoregon.edu

Running Title: FIRST GRADE NUMBER COMPARISON CONGRUENCY

Keywords: dot comparison, nonsymbolic number comparison, executive function, congruency effect, math development

Full reference: Wilkey, E. D., Shanley, L., Sabb, F., Ansari, D., Cohen, J. C., Men, V., Heller, N. A., & Clarke, B. (2021). Sharpening, focusing, and developing: A study of change in nonsymbolic number comparison skills and math achievement in 1st grade. Developmental Science, 25(3), e13194. https://doi.org/10.1111/desc.13194

Publication Date: 11 Dec 2021, First published: 20 November 2021

Data availability statement: The data and analysis code that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Funding statement: This work is supported by the National Science Foundation (NSF) grants DRL 1660840 and DRL 1748954 to Drs. Ben Clarke, Hank Fien, Fred Sabb, and Lina Shanley and by the U.S. Department of Education (USDE) Institute for Education Sciences (IES) grant R324A160046 to Drs. Ben Clarke, Christian Doabler, and Hank Fiend at the University of Oregon. Eric D. Wilkey is the recipient of a Banting Postdoctoral Fellowship (NSERC) and BrainsCAN Postdoctoral Fellowship at Western University, funded by the Canada First Research Excellence Fund (CFREF).

Conflict of interest disclosure: The authors have declared that no competing interests exist.

Ethics approval statement: This study was conducted in accordance of the APA ethical standards and we received approval from the University of Oregon institutional review board.

Data Links for Review Process:

Our current IRB approvals allow for sharing data upon reasonable request and we understand the journal to be requesting data and code availability for the peer review process. Data and analysis code have been archived on the Open Science Framework and are available for access during peer review at the following link. These same files will be archived here for fulfilling data requests in the future.

https://osf.io/d2vzs/?view_only=8955e4c6563c454d848719405f91c618

Research Highlights

- Children's ability to discriminate nonsymbolic number improves throughout development. Competing theories suggest improvement due to sharpening magnitude representations or changes in attention and inhibition.

- The current study intestigates change in nonsymbolic number comparison performance during first grade and whether symbolic number skills, math skills, or executive function predict change.

- Children's performance increased across visual control conditions (i.e. congruent or incongruent with number) suggesting an overall sharpening of number processing.

- Symbolic number skills predicted change in nonsymbolic number comparison performance.

Abstract

Children's ability to discriminate nonsymbolic number (e.g. the number of items in a set) is a commonly studied predictor of later math skills. Number discrimination improves throughout development, but what drives this improvement is unclear. Competing theories suggest it may be due to a sharpening numerical representation or an improved ability to pay attention to number and filter out non-numerical information. We investigate this issue by studying change in children's performance (N = 65) on a nonsymbolic number comparison task, where children decide which of two dot arrays has more dots, from the middle to the end of 1st grade (Mean age at Time 1 = 6.85 years old). In this task, visual properties of the dots arrays such as surface area are either congruent (the more numerous array has more surface area) or incongruent. Children rely more on executive functions during incongruent trials, so improvements in each congruency condition provide information about the underlying cognitive mechanisms. We found that accuracy rates increased similarly for both conditions, indicating a sharpening sense of numerical magnitude, not simply improved attention to the numerical task dimension. Symbolic number skills predicted change in congruent trials, but executive function did not predict change in either condition. No factor predicted change in math achievement. Together, these findings suggest that nonsymbolic number processing undergoes development related to existing symbolic number skills, development that appears not to be driving math gains during this period.

1. Introduction

Like many other animals, humans demonstrate the ability to perceive numerical information early in development. For example, infants can notice the difference between two dots sets of small numbers that differ by a factor of 3 (Smyth & Ansari, 2020). The cognitive system used to process this nonsymbolic numerical information is often referred to as the approximate number system (ANS). This system has been studied closely for over 20 years in large part because individual differences in the ANS are known to influence mathematics development (ANS; Dehaene, 1997; Feigenson et al., 2004), an academic skill that has wide-ranging impacts for future life outcomes (Duncan et al., 2007; Hibbard et al., 2007). The most common experimental task used to index the acuity of the ANS in research supporting this finding is the nonsymbolic number comparison task. In this task, a participant chooses which of two groups of objects (e.g. dots or squares) is greater in number. Many studies and meta-analyses have shown that performance on number comparison tasks correlates with math achievement (Chen & Li, 2014; Fazio et al., 2014; Schneider, Beeres, Coban, Merz, Susan Schmidt, Stricker, & De Smedt, 2017), and further, that individuals with math learning deficits perform very poorly in this task (Mazzocco et al., 2011; Piazza et al., 2010; Price et al., 2007). Given these findings, nonsymbolic number skill has been suggested as a useful component of early screening for math learning difficulties (Butterworth, 2012; Geary et al., 2009; Nosworthy et al., 2013) and as a target for early intervention (Park & Brannon, 2013, 2014; Szűcs & Myers, 2017).

However, a body of recent work questions whether processing of numerical magnitudes is driving the relation between number comparison performance and mathematics. Instead, the task may be confounded by executive function demands that are engaged when resolving conflict between competing aspects of the numerical stimuli. Specifically, stimuli composed of dot sets are generated with visual cues (e.g. surface area, dot sizes, cumulative dot perimeter, density etc.) that are either congruent or incongruent with the numerosity of the dot sets. For example, in some trials, the more numerous dot array would have a greater total surface area (i.e. congruent trials), while in other trials the more numerous dot array would have a smaller total surface area (i.e. incongruent trials). This visual control forces participants to attend to numerosity rather than rely on visual cues that covary with number. Several studies have demonstrated that only performance on incongruent trials correlates with math achievement (Fuhs & McNeil, 2013 - preschool; Gilmore et al., 2013 - ages 4-12; Wilkey et al., 2018- 3rd and 4th grade children) and that children with math deficits only differ from typically developing peers on incongruent trial performance (Bugden & Ansari, 2015 - ages 9-13; Wilkey et al., 2018 - 6th grade), even when controlling for individual differences in executive function in non-numerical tasks. This unique relation between incongruent trials and math achievement, even after controlling for domain-general EF, has led multiple research groups to suggest an important role for number-specific inhibition or attention to number (Fuhs et al., 2016; Piazza et al., 2018; Wilkey et al., 2018; Wilkey & Price, 2018). These findings indicate an important role for the interaction between magnitude perception and executive function in the development of math skills across a wide age range and raise several questions about their development which are the focus of the current study.

1.1 Does numerical perception improve via a sharpening ANS or better attention to number?

It is well known that children become increasingly accurate in processing numerosity with both age and education (Halberda et al., 2012; Halberda & Feigenson, 2008; Landerl & Kölle, 2009; Odic, 2018; Odic et al., 2013). Thus far, increase in performance in the number comparison task has mostly been interpreted as evidence of developmental increases in the acuity, or precision, of the mental representation of number (*i.e., sharpening hypothesis*). However, given the body of recent work, alternative hypotheses have been suggested that rely more on the development of children's ability to attend to, or focus on, number (e.g. *filtering hypothesis* or *attention to number*, (Piazza et al., 2018; Wilkey et al., 2018; Wilkey & Price, 2018). Re-analysis of cross-sectional analyses comparing children aged 3-6, 8-12, and adult support the filtering hypothesis over the sharpening hypothesis by demonstrating a growth in children's ability to focus on numerical properties of stimuli (Piazza et al., 2018). Similarly, ANS training studies have shown that children's performance increased only on number comparison trials with incongruent visual cues, which also suggests that children are improving at filtering out irrelevant information (Fuhs et al., 2016). In the case of a sharpening of the ANS, one would expect an increase in performance across both congruent and incongruent trials. Additionally, these two sources of development may work in concert and are not mutually exclusive. However, as of yet, no study has investigated this question with a longitudinal study in the absence of a targeted intervention that may bias task-specific changes. Therefore, our first research question addresses how performance on the nonsymbolic number comparison task changes over time. We analyze task performance with respect to visual cues that are either congruent or incongruent with stimulus numerosity in order to understand if numerical perception improves as a function of the *sharpening* or *filtering* process in a sample of 1st grade students.

1.2 What influences the development of numerical perception?

As most research investigating the perception of numerical magnitude is ultimately concerned with identifying when and how to intervene to improve numerical skills, the natural next question is—*What influences change in numerical perception?* Increased acuity of numerical magnitude perception has been associated with the acquisition of symbolic number knowledge (Matejko & Ansari, 2016; Mussolin et al., 2014), formal math instruction (Lyons et al., 2018; Piazza et al., 2013; Suárez-Pellicioni & Booth, 2018), and the development of executive functions (Fuhs et al., 2016; Gilmore et al., 2013). However, increased acuity due to *sharpening* or *filtering* may be differentially affected by these other factors. For example, it may be that children's ability to filter out irrelevant cues develops as a domain-general ability to filter out any irrelevant information. On the other hand, sharpening may occur as children acquire the use of exact, symbolic numerical values. While nonsymbolic and symbolic number skills are often correlated, their interdependent development has not been clearly articulated, especially with reference to the concurrent development of executive function skills. Therefore, our second set of study questions focus on what influences the change in accuracy rate on congruent and incongruent trials of the number comparison task, including math achievement, executive function, and symbolic number processing skills.

1.3 What influences the relation between numerical magnitude perception and math achievement?

Lastly, if the development of numerical perception is influenced by multiple cognitive mechanisms, which mechanisms most closely relates to mathematical skills? Nonsymbolic number comparison performance in early childhood has been shown to correlate with math achievement even when considering the influence of non-numerical visual parameters of task stimuli and inhibitory control (Keller & Libertus, 2015; Starr et al., 2017). However, other research shows that the relation between number comparison performance and math is either partially (Gilmore et al., 2015; Keller & Libertus, 2015; Wilkey et al., 2018) or completely explained by individual differences in non-numerical executive function (Fuhs & McNeil, 2013; Gilmore et al., 2013). So, while it is well established in the meta-analytic literature that there is a small to medium effect size in the relation between nonsymbolic numerical magnitude perception and mathematical ability (*r* = .241, *k* = 195; (Schneider, Beeres, Coban, Merz, Susan Schmidt, Stricker, De Smedt, et al., 2017)), the specific factors that drive this relation are not well understood. It may be that performance on incongruent trials or congruent trials is differentially related to growth in math skills as a function of individual differences in symbolic number development or executive function abilities. To address this issue, our last study question investigates what factors predict math achievement alongside nonsymbolic discrimination.

1.4 The Current Study

This study aims to address (a) whether improvement in nonsymbolic number skills is due to a *sharpening* of magnitude representations or the developing ability to focus on number (i.e. *filtering*), (b) what cognitive mechanisms influence this change, and (c) what these factors may tell us about the relation between nonsymbolic number skills and math skills. To do this, we focus our analyses on change in performance on a nonsymbolic number comparison task independently for trials with congruent and incongruent visual cues in a sample of children measured at the middle and end of 1st grade. First, we investigate change in accuracy over time in the nonsymbolic number comparison task and then conduct a series of moderator and mediator analyses of that change related to symbolic number skills and executive function. Lastly, we explore the influence that potential moderating factors have in the relation between nonsymbolic number comparison performance and math achievement, split by congruency. Analyses were preregistered: https://osf.io/v2uq9/?view_only=73c21ac7cd0d42d8b8be55786c54f7fe

2. Method

2.1 Participants

The analytic sample for the current study is comprised of students in the first three cohorts of a multi-year National Science Foundation (NSF; DRL 1748954 & DRL 1660840) funded study aimed at examining cognitive and neural correlates of first grade mathematics development. Year 1 and year 2 participants were recruited from schools that participated in a large-scale efficacy trial of a first grade mathematics intervention funded by the Institute of Education Sciences (IES; R324A160046). Year 3 participants were recruited from schools who continued to implement first grade mathematics intervention after the conclusion of the IES study. While the primary aim of the IES study was to investigate the efficacy of an evidenced-based mathematics intervention for students at risk for mathematics difficulties, students of all mathematics abilities were recruited for the NSF study. In all 121, students participated in years 1-3 of this study. In the full sample, 55% reported their biological sex as male. Additionally, 1% of participants identified as Asian, 1% identified as Black, 4% identified as Native Hawaiian/Pacific Islander, 94% identified as White, 11% identified as Hispanic or Latino, and 6% were reported as more than one race. Of these students, 12% were eligible for special education and 5% met criteria for limited proficiency in English. District, school, and classroom data are presented in **Appendix A**. First grade is an ideal time to study change in both symbolic and nonsymbolic number representations since children are paying more explicit attention to number during formal math instruction. From an assessment standpoint, many children with math learning difficulties demonstrate for the first time that they are lagging behind their peers in math skill acquisition. Therefore, first grade represents the earliest opportunity to intervene formally and to assess math skills across a sample that captures response to formal schooling.

2.2 Analytic Sample

The analytic sample for the current study consists of all children for whom we had complete data including: (1) nonsymbolic number comparison at T1 and T2, (2) symbolic number comparison at T1 and T2, (3) math achievement at T1 and T2, (4) Head, Toes, Knees, and Shoulders (HTKS) for at least one time point, and (5) oral reading fluency for at least one time point. One child was excluded because they received a score of 0 on the Head, Toes, Knees, and Shoulders task, indicating they did not understand the task. The resulting final analytic sample included 65 participants. See **Table 1** for demographic information for the analytic sample. Descriptive statistics of study measures are presented in **Table 2**. Bivariate correlations are presented in **Table 3**.

Table 1. Descriptive Statistics for Student Characteristics (*n* = 65)

Student characteristic	n (%)
Male	35 (54%)
Race/Ethnicity	
American Indian or Alaska Native	0 (0%)
Asian	0 (0%)
Black	0 (0%)
Native Hawaiian/ Pacific Islander	3 (5%)
White	52 (80%)
More than one race	4 (6%)
Hispanic or Latino	6 (9%)
Limited English proficiency	3 (5%)
SPED eligible	6 (9%)
Age at T1, M (SD)	6.85 (0.36)

Note. Mean and standard deviation reported for age. SPED eligible = students who are eligible to receive special education services based on a qualifying disability.

Table 2. Descriptive Statistics for All Measures (N = 65)

	Mean	**Median**	**Min**	**Max**	**SD**	**Possible**
NS Comparison (T1)	80.4	82.4	53.5	96.8	11.00	101
NS Comparison (T2)	84.1	86.5	57.5	98.2	9.27	101
NS Comparison Congruent (T1)	81.9	87.0	46.7	98.9	12.4	101
NS Comparison Incongruent (T1)	78.9	82.4	51.9	101.0	12.0	101
NS Comparison Congruent (T2)	85.3	87.0	56.1	96.0	8.72	101
NS Comparison Incongruent (T2)	82.9	84.2	53.3	101.0	11.4	101
ASPENS: Symbolic MC (T1)	15.9	15	0	37	7.62	-
ASPENS: Symbolic MC (T2)	21.2	21	3	35	6.57	-
ASPENS: BF(T1)	4.5	3	0	13	3.57	-
ASPENS: BF (T2)	8.8	8	0	26	5.95	-
TEMA-3 (T2)	47.9	47	33	71	9.42	72
Heads, Toes, Knees, and Shoulders	46.2	49	20	60	10.20	94
Oral Reading Fluency	57.0	45	6	163	41.5	-

Note. T1 = Time 1; T2 = Time 2; NS Comparison = nonsymbolic number comparison; Symbolic MC = Magnitude Comparison subtest of ASPENS; BF = Basic Arithmetic Facts and Base 10 subtest of ASPENS; TEMA-3 = Test of Early Mathematics Achievement, 3rd Edition. ASPENS measures and Oral Reading Fluency do not have a maximum possible score.

Table 3. Bivariate Correlations Between All Study Measures (n = 65)

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1. NS Comparison CON (T1)	—									
2. NS Comparison INC (T1)	0.573***	—								
3. NS Comparison CON (T2)	0.302*	0.413***	—							
4. NS Comparison INC (T2)	0.328**	0.365**	0.669***	—						
5. ASPENS: Symbolic MC (T1)	0.288*	0.255*	0.296*	0.285*	—					
6. ASPENS: Symbolic MC (T2)	0.171	0.280*	0.288*	0.225	0.782***	—				
7. ASPENS: BF (T1)	0.292*	0.310*	0.294*	0.270*	0.657***	0.579***	—			
8. ASPENS: BF (T2)	0.317*	0.299*	0.179	0.183	0.537***	0.588***	0.607***	—		
9. TEMA-3 (T2)	0.238	0.233	0.265*	0.206	0.641***	0.612***	0.748***	0.531***	—	
10. HTKS	0.300*	0.212	0.018	0.026	0.222	0.097	0.317*	0.300	0.300*	—
11. Oral Reading Fluency	0.239	0.207	0.179	0.141	0.643***	0.480***	0.446***	0.481	0.529***	0.248*

Note. T1 = Time 1; T2 = Time 2; NS Comparison = nonsymbolic number comparison; CON = congruent trials; INC = incongruent trials; Symbolic MC = Magnitude Comparison subtest of ASPENS; BF = Basic Arithmetic Facts and Base 10 subtest of ASPENS; HTKS = Head, toes, knees, shoulders.
* *p* < .05, ** *p* < .01, *** *p* < .001

2.3 Power

Power analyses were conducted before analysis but after data collection and documented in the secondary data analysis preregistration in order to address the feasibility of the current data to address the study questions. We calculated a power analysis based on Bugden and Ansari (2016) for the most critical parts of the current analysis. Most of the central questions in the current analysis depend on the effect of congruency of visual cues of the number comparison task. Bugden and Ansari report a congruency effect with an effect size of Cohen's d = 0.719 across their typically developing and dyscalculic sample in the same Panamath task used in the current study. To account for publication bias, and the small sample in the study by Bugden and Ansari (n = 24), we halved the effect size of Bugden and Ansari (2016) and determined the number of subjects needed to observe a congruency effect using the pwr toolbox in R (Champely, 2020). In order to have power = 0.8 to detect an effect of congruency in a paired samples t-test, we would need a sample of n = 63.

We also calculated the number of participants needed to detect a correlation between performance in Panamath and our outcome of interest, the TEMA-3. Schneider et al. (2017) estimated this correlation to be r = 0.413. Given that this correlation is based on k = 37 effect sizes in a meta-analysis, which was checked for (and did not indicate evidence of) publication bias, we did not halve the effect size as above. In order to have power = 0.8 to detect a correlation between performance in the number comparison task and math achievement as measured by the TEMA-3, we would need a sample of n = 43.

2.4 Procedure

Participants were recruited via letters distributed to their families by school administrators and classroom teachers. Participants were briefly screened via email or phone and then scheduled for a research appointment. Because the broader project involved an MRI component, children who had non-removable metal devices (e.g. braces, hearing aids) were excluded from the study. Panamath and reading assessment activities were conducted in conjunction with MRI research appointments at the Lewis Center for Neuroimaging (LCNI) at the University of Oregon. Other academic and behavioral measures were collected by research assistants (RAs) at participating schools in one, one-on-one session unless scheduling contraints required the session to be conducted across two days. School-based data collection activities were completed prior to scheduling research appointments in the lab.

After reviewing the parent informed consent and obtaining child assent in a private testing room, students first completed the Dynamic Indicators of Basic Early Literacy Skills 6th edition (DIBELS 6th edition, see description below) then MRI acclimation and scanning activities. After scanning, a nonsymbolic number comparison assessment (i.e., Panamath, described below) was completed on a computer with a trained project research assistant (RA) in a private testing room. Throughout the research appointment, RAs supervised all sessions to monitor completion of required tasks, family satisfaction, and safety of research activities. Participants who successfully completed all research activities in their initial appointment (i.e., T1) were invited back for a second research visit approximately 4-5 months later (i.e., T2). Average time between the T1 and T2 lab visit was 4.5 months (range = 3.1 – 6.1).

2.5 Measures

2.5.1 Nonsymbolic Number Comparison

ANS acuity was assessed using the Panamath version 1.22 software (Halberda et al., 2008). PanaMath is a free-standing software (see http://panamath.org) suitable for administration to subjects ranging from 3-85 years. This assessment measures approximate number system

(ANS) aptitude by prompting participants to “determine which color has more dots” based on “a flash of colored dots on the screen.” Ratios are presented for 1,951 ms and participants are prompted to press “F” for more yellow dots or “J” for more blue dots, then space bar to advance. There are no practice trials and item feedback is not given. Occasional praise of effort to encourage persistence was employed, as needed. Participants are informed that the “experiment consists of many trials,” the display includes a progress bar, and they are informed that they can end the experiment early by pressing the “esc” button, but they are not aware of timing or age-related item presets.

Prior to administration, subject age, ID, and an estimated administration time of 5 minutes was entered in the administrative interface. Depending on each individual’s speed of response, the number of trials varied by participant. While most participants completed the assigned 72 trials, one participant at each timepoint completed fewer trials due to fatigue. Due to an undetermined technical issue, 3 children at T1 and 1 child at T2 received more than 72 trials. At Time 1, the total number of trials completed by participants ranged from 26 to 160 trials (Mean = 73.5). At Time 2, the total number of trials completed by participants ranged from 45 to 80 (Mean = 71.4). Ratios presented ranged from 1.34 to 2.94 and dots presented ranged from 5 dots to 21 dots. In approximately half of the trials, the surface area of the dots were proportional to the number of dots within the array and the average dot size was equated (dot-size controlled; coded 0 in Panamath). In these trials, the total surface area of the dots indicated the more numerous dot array, which we refer to as congruent trials. In the other half of the trials, the total surface area was equal between the two dot arrays (area controlled, coded -1 in Panamath). In these trials, the size of the individual dots was negatively correlated with numerosity, referred to as incongruent trials. For these trials, children could not select the larger dot array by relying on the amount of color occupying space on the computer screen and dot size provided an incongruent visual cue (see **Figure 1**). Accuracy rates were calculated separately for congruent and incongruent trials. Mean accuracy rates were adjusted in order to equate task version difficulty across timepoints and participants (see **Appendix B** for a discussion of dependent variable selection and full details of the adjustment).

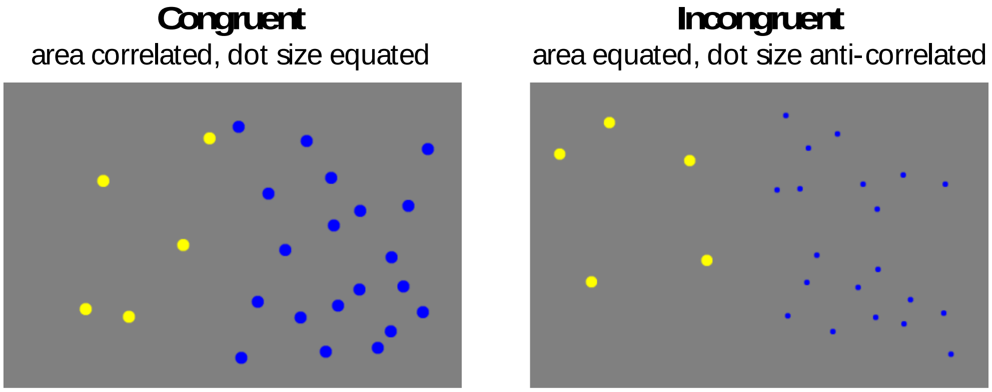
Congruent
area correlated, dot size equated
Incongruent
area equated, dot size anti-correlated

Figure 1. Example of congruent and incongruent stimuli administered in the Panamath task for a ratio of 2.0, or 20 dots versus 5 dots.

Shapiro-Wilk tests of normality indicated that accuracy were not normally distributed within timepoints and congruency conditions [all *p* < .05, skewness for congruent T1 = -1.02, incongruent T1 = -0.492, congruent T2 = -0.932, incongruent T2 = -0.711]. Therefore, accuracy rate scores were transformed by first reflecting them (subtracting each score from the maximum value across all participants plus 1) and then taking the square root. Transformation reduced skewness to levels we deemed acceptable, but the Shapiro-Wilk test was still significant for congruent trials at Time 2 [skewness for congruent T1 = 0.381, *p* = .100; incongruent T1 = -0.173, *p* = .278; congruent T2 = 0.271, *p* = .016; incongruent T2 = 0.063, *p* = .687]. Raw scores are reported for descriptive statistics, but transformed scores are used for all analyses. Transformed scores were reversed for analyses to maintain a higher is better coding scheme.

2.5.2 Executive Function

The Heads, Toes, Knees, and Shoulders task (HTKS) is an observational assessment of behavioral self-regulation that measures a child's ability to inhibit imitative responses, focus and shift attention, and remember and apply multiple rules. The HTKS takes approximately five minutes to complete and participants receive two points for correctly responding to prompts on the first attempts, one point for items with self-corrections, and zero points for incorrect responses. After a brief 4-item introductory practice phase, HTKS contains three parts each with training and practice phases that allow for feedback and a test phase where no feedback is given. Particpants only advance to latter parts if they meet performance criteria on the previous part. The first part contains a total of 16 items, 6 of which (2 in training and 4 in practice) allow for up to 3 corrections from the test administrator. Participants must achieve a score of 4 or higher to advance to the next part. Part 2 contains a total of 15 items, 5 of which (1 training and 4 practice) allow for up to 3 corrections from the test administrator. Participants must achieve a score of 4 or higher on part 2 to advance to the final part. Section 3 contains a total of 16 items, 6 of which (2 training and 4 practice) allow for up to 2 corrections from the test administrator. Across all three sections, there are a total of 94 points possible. Interrater reliability for the task is high (.95; Ponitz et al., 2008). The HTKS is positively correlated with (a) parent ratings of attentional focusing (r = .25) and inhibitory control (r = .20), and (b) teacher ratings of classroom behavioral regulation (r = .20). Further, HTKS administered in the fall of kindergarten was a significant predictor of spring math performance (d =.56). HTKS is the raw number of items correct (Cameron Ponitz et al., 2008; Ponitz et al., 2009).

A Shapiro-Wilk test of normality indicated that raw scores were not normally distributed [p < .001, skewness = -.882], so the scores were transformed by first reflecting them (subtracting each score from the maximum value across all participants plus 1) and the taking the square root

[Shapiro-Wilk *p* = .208, skewness = 0.259]. Again, raw scores are reported for descriptive statistics, but transformed scores are used for all analyses. Transformed scores were reversed for analyses to maintain a higher is better coding scheme.

2.5.3 ASPENS: Symbolic Magnitude Comparison

Assessing Student Proficiency in Early Number Sense (ASPENS; (Clarke et al., 2011) is a series of four brief (1- to 2-min) measures designed to assess student understanding of critical number concepts. The measure assesses four early math skills: (a) numeral identification, (b) magnitude comparison, (c) missing number, and (d) basic arithmetic facts and base 10. ASPENS measures are timed and individually-administered, and all subtests have a discontinue criteria of 5 consecutive incorrect answers. While only the magnitude comparison and basic arithmetic facts and base 10 subtests scores were used in the current study, all subtests are described in more detail below. In the *numeral identification* subtest, participants complete 2 practice items and then are presented with a list of numerals ranging from 0 – 20 and prompted to move across the page starting at the top of the page and name as many numbers as they can. Participants receive one point for every numeral correctly identified in 1 minute. In *the magnitude comparison* subtest (hereafter ASPENS: Symbolic MC), participants are shown two numbers (randomly sampled from 0-99 in 1st Grade) presented side-by-side in a box and prompted to verbally indicate the larger number. Two practice items are presented with feedback and a participant's score is the number of items answered correctly in 1 minute. Simlarly, in the *missing number* subtest, after two practice items with feedback, participants are shown boxes containing two numbers and a blank (placed in either the beginning, __ 12 13; middle, 76 __ 78; or end of the sequence, 1 2 __) and prompted move across the page to verbally identify the number that goes in each blank. Participants receive one point for every correctly identified missing number in 1 minute. Finally, in the *basic arithmetic facts and base 10* subtest, participants are given a pencil and a two-sided worksheet containing twenty addition and subtraction problems within 20 on each side and prompted to work across the page and complete as many problems as they can. Participants receive one point for every correct answer in 2 minutes. Test authors report test-retest reliability ranges from the .70s to .90 across the four subtests. Criterion concurrent validity with the TerraNova 3 is reported as ranging from .51 to .63. Raw scores were used. They were normally distributed [Time 1 Shapiro-Wilk *p* = .734, skewness = .202; Time 2 Shapiro-Wilk *p* = .819, skewness = -.077].

2.5.4 Math Achievement

Math achievement was indexed using two different measures. The first measure of math achievement was the Test of Early Mathematics Ability – 3rd Ed. (TEMA-3; (Ginsburg, H. & Baroody, 2003) which was administered at both Time 1 and Time 2. The TEMA-3 is a standardized measure of informal and formal number and operations knowledge that is widely used in studies of early math intervention. The TEMA-3 is designed for students ages 3 to 8 years 11 months. The TEMA-3 is designed to identify student strengths and weaknesses in specific areas of mathematics, including skills related to counting, number facts and calculations, and related mathematical concepts. Test authors report alternate-form reliability of .97 and test-retest reliability ranges from .82 to .93. Concurrent validity with other criterion measures of mathematics is reported as ranging from .54 to .91.

Our preregistered analyses planned to use the TEMA-3 scores for math achievement, however, a substantial number of children were missing scores for the TEMA-3 at Time 1 (n = 16 of 65, or 25%). Since the full analysis requires a math achievement score for Time 1 and Time 2, scores from the ASPENS measure Basic Arithmetic Facts and Base 10 (hereafter Basic

Facts) were also used as a math achievement measure to control for math achievement at Time 1. As described above, in the ASPENS Basic Facts subtest, participants are provided a set of written basic facts problems to solve including addition and subtraction problems. Participants are instructed to work left-to-right and top-to-bottom to complete the problems and given two minutes to solve as many problems as possible. Supplementary analysis are included comparing ASPENS: Basic Facts and TEMA-3 for the math achievement analyses.

Both TEMA-3 and ASPENS: Basic Facts scores were positively skewed [Time 2 TEMA-3 Shapiro-Wilk p = .012, skewness = .642; ASPENS Basic Facts Time 1 Shapiro-Wilk p < .001, skewness = 0.814; ASPENS Basic Facts Time 2 Shapiro-Wilk p < .001, skewness = 0.950]. Square-root transformed scores are used for all analyses [transformed scores: Time 2 TEMA-3 Shapiro-Wilk p = .071, skewness = .444; ASPENS Basic Facts Time 1 Shapiro-Wilk p = .055, skewness = -0.081; ASPENS Basic Facts Time 2 Shapiro-Wilk p = .515, skewness = -0.099] but raw scores are reported for descriptive statistics.

2.5.5 Reading Fluency

In the current study, we included a measure of reading fluency to use as a control measure when predicting math achievement and growth in math achievement. Since reading fluency and math achievement are typically correlated and increase as general academic knowledge increases, controlling for reading fluency results in more domain-specific results. The Dynamic Indicators of Basic Early Literacy Skills (DIBELS; (Good & Kaminski, 2002) Oral Reading Fluency subtest (ORF) was used as a measure of reading fluency. The DIBELS: ORF is a standardized, individually administered test of accuracy and fluency with connected text. Student performance is measured by having students read a passage aloud for one minute. The number of correct words per minute is the oral reading fluency score. In this study, students completed three brief ORF passages and the median raw score was retained. A Shapiro-Wilk test of normality indicated that raw scores were not normally distributed [$p < .001$, skewness = 0.824], so the scores were square-root transformed [Shapiro-Wilk $p = .059$, skewness = 0.266]. Again, raw scores are reported for descriptive statistics, but transformed scores are used for all analyses.

2.6 Analysis and Software

Analyses were conducted using a mixture of R (Team, 2018; Wickham, 2017) and jamovi (*The Jamovi Project*, 2019). ANOVAs and regression analyses were conducted in jamovi and hand-checked in R (Fox & Weisberg, 2018). Plots were created using the "ggplot2" package in R (Wickham, 2016). Mediation was conducted using the "medmod" and ""jAMM" packages implemented in jamovi, using the defaults settings and the "standard" Delta method for calculating confidence intervals. Both packages estimate mediation coefficients using Maximum Likelihood method implemented in lavaan R package (Rosseel, 2012).

3. Results

3.1 Change in Number Comparison Over Time: Evidence for Sharpening or Filtering

To address our first research question related to the development of numerical magnitude perception, we conducted a repeated measures, two-way ANOVA to assess the main effects of Congruency and Time, and their interaction. We reasoned that if the sharpening hypothesis is supported by the data, accuracy rates would increase on both congruent and incongruent trials, since number processing is involved in both conditions. If the filtering hypothesis was supported, we would see increased accuracy mainly on the incongruent trials where inhibition and selective attention are more heavily taxed, resulting in an interaction demonstrating a greater improvement for incongruent trials. Despite these contrasting hypotheses, both sharpening and filtering may be a simultaneous source of improved accuracy rates and are not mutually exclusive. It was possible that we would see increased improvement for both conditions, with a bigger effect for incongruent trials, indicating both increased precision and increased filtering ability.

This analysis revealed a main effect of Congruency [$F(1, 64) = 43.67$, $p < .001$, $\eta^2 = 0.406$], a main effect of Time [$F(1, 64) = 8.26$, $p = .006$, $\eta^2 = 0.114$], an no Congruency x Time interaction [$F(1, 64) = 0.01$, $p = .917$, $\eta^2 = 0.000$](see **Figure 2** for means). On average, children were 2.71 points [95% CI: 1.11 – 4.31] more accurate for Congruent trials than Incongruent trials [$t(64) = 6.61$, $p < .001$, Cohen’s $d = 0.820$] and were 3.68 points [95% CI: 0.99 – 6.38] more accurate at Time 2 compared to Time 1 [$t(64) = 2.87$, $p = .006$, Cohen’s $d = 0.356$]. The simple effects for change over time within congruency condition were also significant. For Congruent trials, accuracy was 3.40 points higher [95% CI: 0.282 – 6.53] at Time 2 than at Time 1 [$t(64) = 2.27$, $p = .026$, Cohen’s $d = 0.282$]. For Incongruent trials, accuracy was 3.96 points higher [95% CI: 0.68 – 7.24] at Time 2 than at Time 1 [$t(64) = 2.49$, $p = .015$, Cohen’s $d = 0.309$]. Further, the simple effects of congruency within timepoint were significant. At Time 1, accuracy was 2.99 points higher [95% CI: 0.37 – 5.62] for Congruent than Incongruent trials [*t*(64) = 4.13, *p* < .001 , Cohen's *d* = 0.512]. At Time 2, accuracy was 2.44 points higher [95% CI: 0.38 – 4.49] for Congruent than Incongruent trials [*t*(64) = 4.94, *p* < .001 , Cohen's *d* = 0.612].

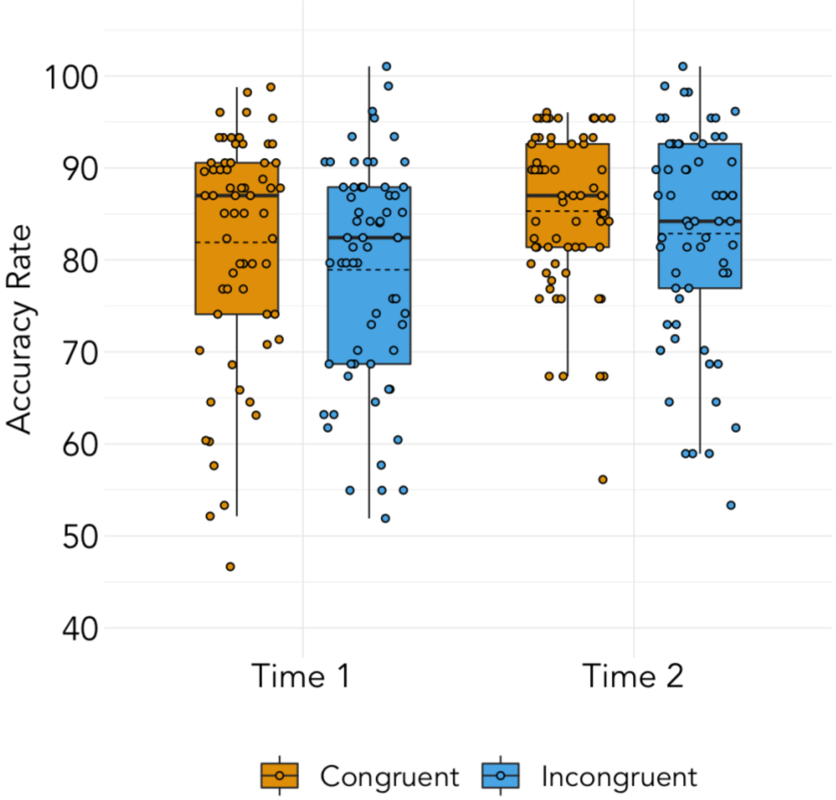

Figure 2. Nonsymbolic number comparison accuracy rates by Time and Congruency condition. Congruent trials are labeled in orange and Incongruent trials in blue. Box plot hinges represent 25th and 75th percentile of distributions, whiskers extend from hinge to the largest value not beyond 1.5 times the interquartile range, the middle solid line represents the median value, and the middle dashed line represents the mean.

Most directly related to the study question, results indicate that children's performance increased over time but that the rate of increase did not differ by congruency. Therefore, in the current sample, it appears that children's accuracy is increasing due to a general sharpening of magnitude perception rather than enhanced filtering of non-numerical information that would lead to increased performance mostly on the incongruent trials.

3.2 Factors Moderating and Mediating Change in Performance Over Time

We next performed a series of analyses to better understand what other cognitive factors may influence growth in nonsymbolic number comparison performance as a factor of congruency. For example, whereas symbolic number skills may be a predictor of growth in Congruent trials as a critical component of sharpening one's sense of magnitude, executive function may be more influential for growth in Incongruent trials. Specifically, we first performed a series of moderator analyses via regression models and then investigated whether growth in symbolic number skills mediated the growth in nonsymbolic number comparison for each congruency condition.

3.2.1 Executive Function, Symbolic Number Skills, and Math Achievement as Moderators

To investigate what factors moderate the increase in performance in the nonsymbolic number comparison task from Time 1 to Time 2, we ran regression models that predict accuracy rate in the number comparison task for Congruent trials at Time 2 (**Table 4**) and Incongruent trials at Time 2 (**Table 5**). Three models are shown for the three moderators of interest: math achievement (Model 1 (M1) - ASPENS: Basic Facts), executive function (Model 2 (M2) – Head, Toes, Knees, and Shoulders (HTKS)), and symbolic number skills (Model 3 (M3) – ASPENS Magnitude Comparison (MC)). In each regression, accuracy rate for the respective congruency condition at Time 1 is included as the first predictor in order to control for performance at Time 1. Accordingly, the dependent variable should be interpreted as growth in accuracy rate between the two time points. The moderator of interest is the second term and indicates whether the variable predicts change in accuracy rate. The third term is the interaction of the moderator of interest and the accuracy rate for the number comparison task at Time 1.

Results for the growth in accuracy rate for Congruent trials (**Table 4**) indicate that math achievement and executive function did not predict change, but that symbolic number skills, as measured by the ASPENS Magnitude Comparison task, did predict change in accuracy rate such that participants with higher initial symbolic number skills were predicted to demonstrate greater growth in accuracy on Congruent trials [standardized β (3, 61) = 0.944, p = .027]. And, while the interaction term for ASPENS Magnitude Comparison was approaching significance [standardized β = -0.977, p = .078], it was not a statistically significant predictor. In sum, only symbolic number skills was associated with the growth of accuracy rate on congruent trials of the nonsymbolic number comparison task.

Table 4. Moderator Analysis for Growth in Congruent Trials of Nonsymbolic Number Comparison

Predictor	M1	M2	M3
NNC Congruent (T1)	0.457 [-.09 – 1.00]	0.725* [0.07 – 1.38]	0.724* [0.13 – 1.32]
ASPENS: BF (T1)	0.545 [-.20 – 1.29]		
NNC Congruent (T1) x ASPENS: BF (T1)	-0.452 [-1.45 – 0.54]		
HTKS		0.377 [-0.36 – 1.12]	
NNC Congruent (T1) x HTKS		-0.709 [-1.79 – 0.37]	
ASPENS: Symbolic MC (T1)			0.944* [0.11 – 1.78]
NNC Congruent (T1) x ASPENS: Symbolic MC			-0.997 [-2.11 – 0.11]
R^2	0.149*	.122*	0.182**

Note. Regression coefficients are standardized. 95% confidence intervals are in brackets. NNC = nonsymbolic number comparison; BF = Basic Arithmetic Facts and Base 10 subtest of ASPENS; HTKS = Head, toes, knees, shoulders. ASPENS: Symbolic MC = Magnitude Comparison subtest of ASPENS.
**p* < .05, ***p* < .01,

Table 5. Moderator Analysis for Growth in Incongruent Trials of Nonsymbolic Number Comparison

Predictor	**M1**	**M2**	**M3**
NNC Incongruent (T1)	0.338 [-.17 – 0.85]	0.769 [-0.03 – 1.57]	0.334 [-0.31 – 0.98]
ASPENS: BF (T1)	0.207 [-.40 – 0.81]		
NNC Incongruent (T1) x ASPENS: BF (T1)	-0.051 [-0.90 – 0.79]		
HTKS		0.299 [-0.43 – 1.03]	
NNC Incongruent (T1) x HTKS		-0.593 [-1.75 – 0.56]	
ASPENS: Symbolic MC (T1)			0.232 [-0.57 – 1.03]
NNC Incongruent (T1) x ASPENS: Symbolic MC			-0.039 [-1.15 – 1.07]
R^2	0.161*	.151*	0.173**

Note. Regression coefficients are standardized. 95% confidence intervals are in brackets. NNC = nonsymbolic number comparison; BF = Basic Arithmetic Facts and Base 10 subtest of ASPENS; HTKS = Head, toes, knees, shoulders. ASPENS: Symbolic MC = Magnitude Comparison subtest of ASPENS.
**p* < .05, ***p* < .01,

Here we further detail what these trends demonstrate in the regression model. On average children could answer about 16 Arabic numeral comparisons in 1 minute. On the lower end of the sample, two children were unable to correctly answer prompts and 18 children scored below 10 (i.e. 6s per item). Children in the top quartile answered between 21 and 37 comparisons in 1 minute. Again, regression results indicate that participants with greater Time 1 symbolic number skills grew more from Time 1 to Time 2 in their nonsymbolic skills, specifically for congruent trials (**Table 4, Model 3**). For example, if a child achieved an accuracy rate of 75.0% at Time 1 for congruent trials of the nonsymbolic number comparison task and performed 1 SD *below* the mean on the Symbolic MC, their estimated Time 2 accuracy rate for the nonsymbolic comparison task would be 76.2%. In contrast, for a child with the same accuracy rate at Time 1 on the nonsymbolic task (75.0%) that performed 1SD *above* the mean on the Symbolic MC task, their predicted accuracy rate would be 88.8% (interaction terms held at the mean, for a full plot of model scores, see **Figure 3**).

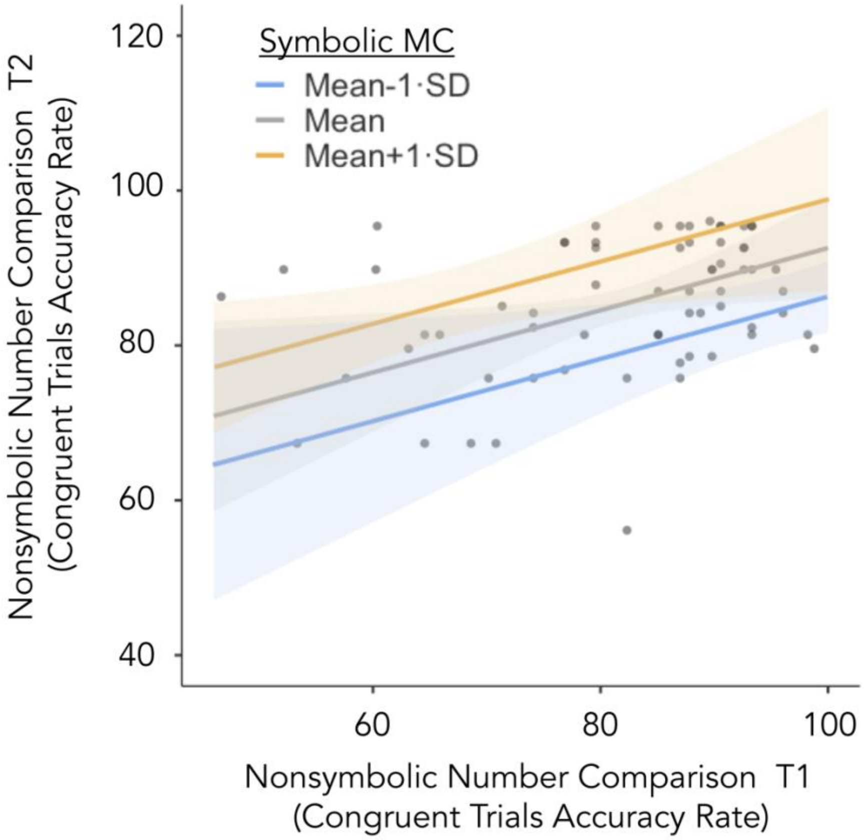

Figure 3. Predicted Time 2 nonsymbolic number comparison accuracy rate as a function of accuracy at Time 1 and ASPENS: Symbolic Magnitude Comparison (Symbolic MC)(Table 4, Model 3), with separate regression lines for the mean performance on Symbolic MC (grey), mean +1 SD (blue), and mean -1 SD. Shaded areas around each regression line indicate the 95% CI; grey dots represent untransformed observed scores. Interaction term is held at the mean across values.

Results for the growth in accuracy rate for Incongruent trials indicate that none of the three potential moderators of interest explained growth in accuracy rate from Time 1 to Time 2 (**Table 5**). Of note, symbolic number skills appears to be a unique predictor of growth in accuracy rate on Congruent trials, since the standardized regression coefficient is notably lower for the Incongruent relation (**β** = 0.232) than the Congruent relation (**β** = 0.944). Although not preregistered, we did explore whether any of the moderators showed differing results when accuracy rates for Congruent and Incongruent trials were combined (see **Supplementary Table**

1). When accuracy rates were combined, none of the three potential moderators was a significant predictor of accuracy rate growth, nor were their interaction terms.

3.2.2 Symbolic Number Skills as Mediator of Growth

To investigate whether the growth in symbolic number skills mediates the growth in performance on the nonsymbolic number comparison task, we conducted mediation analyses for Incongruent and Congruent trials separately, in parallel with the moderator analyses above. Mediation analyses were conducted by examining direct and indirect effects using the following multi-step process: (1) Time 2 accuracy was regressed on Time 1 accuracy, for congruent and in congruent trials in the respective models, (2) Time 1 accuracy was entered as a predictor of change in symbolic comparison skills (i.e. ASPENS: Magnitude Comparison), (3) Time 2 accuracy was regressed on the change in symbolic comparison term, and (4) the indirect effects was requested to examine the extent to which symbolic comparison skills mediated the relation between Time 1 and Time 2 accuracy, for incongruent and congruent trials.

The model indicated that change in symbolic number skills from Time 1 to Time 2 did not mediate the growth of nonsymbolic number comparison accuracy rate for Congruent trials between Time 1 and Time 2. As **Figure 4** (*top*) illustrates, the direct effect between Time 1 accuracy rate and Time 2 accuracy rates was significant [c = 0.253, 95% CI = 0.066 – 0.440, p = .008, standardized β = 0.312]. However, the indirect effect was not significant [a x b = -0.008, 95% CI = -0.037 – 0.020, p = .565, standardized β = 0.066].

Results were similar for the mediation analysis for growth on Incongruent trials. The change in symbolic number skills from Time 1 to Time 2 did not mediate the growth of nonsymbolic number comparison accuracy rate for Incongruent trials between Time 1 and Time 2. As **Figure 4** (*bottom*) illustrates, the direct effect between Time 1 accuracy rate and Time 2 accuracy rates was significant for incongruent trials [c = 0.379, 95% CI = 0.146 – 0.612, p = .001, standardized β = 0.371]. However, the indirect effect was not significant [a x b = -0.006, 95% CI = -0.038 – 0.026, p = .565, standardized β = -0.006].

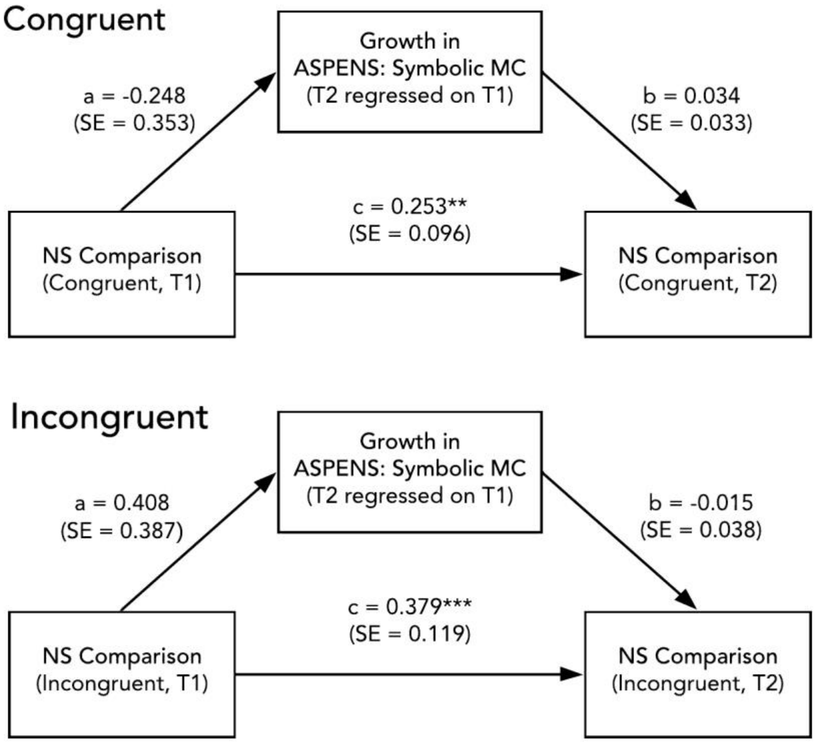
Congruent
Growth in ASPENS: Symbolic MC (T2 regressed on T1)
a = -0.248 (SE = 0.353)
b = 0.034 (SE = 0.033)
c = 0.253** (SE = 0.096)
NS Comparison (Congruent, T1)
NS Comparison (Congruent, T2)
Incongruent
Growth in ASPENS: Symbolic MC (T2 regressed on T1)
a = 0.408 (SE = 0.387)
b = -0.015 (SE = 0.038)
c = 0.379*** (SE = 0.119)
NS Comparison (Incongruent, T1)
NS Comparison (Incongruent, T2)

Figure 4. Mediation models showing the relation between nonsymbolic number comparison (NS Comparison) accuracy rates for (top) congruent trials and (bottom) incongruent trials at Time 1 (T1) and Time 2 (T2) with growth in symbolic number skills (ASPENS: Magnitude Comparison (Symbolic MC), Time 2 regressed on Time 1) as a mediator.
**p < .01, ***p < .001

3.3 Number Comparison Performance and Math Achievement

Our last research question asked whether nonsymbolic number comparison performance predicted math achievement or growth in math achievement in 1st grade, whether those relations were different for Congruent versus Incongruent trials, and how specific those relations were when controlling for other factors, such as symbolic number skills, executive function, and reading fluency. To investigate these relations, we conducted a series of four multiple regressions. The first two multiple regression models predict math achievement at Time 1 from nonsymbolic number comparison accuracy rate for Congruent trials (**Table 6**) and Incongruent trials (**Table 7**). Model 1 in each regression shows the simple linear regression between the two variables. Model 2 adds factors that measure children's symbolic number skills, executive function, and reading fluency in order to determine the predictive nature of nonsymbolic number skills for math achievement while considering a range of other cognitive factors. Results indicate that there is a statistically significant relation between nonsymbolic number comparison accuracy rate at Time 1 and math achievement at Time 1 for both Congruent trials [$\boldsymbol{\beta}$ = 0.292, *p* = .018] and Incongruent trials [$\boldsymbol{\beta}$ = 0.310, *p* = .012]. However, when the additional factor are considered, symbolic number skills (ASPENS: Magnitude Comparison) is the only significant predictor of math achievement. Nonsymbolic number comparison accuracy rate at Time 1 is no longer a significant predictor of math achievement in Model 2 for Congruent trials [$\boldsymbol{\beta}$ = 0.071, *p* = .018] or Incongruent trials [$\boldsymbol{\beta}$ = 0.310, *p* = .012]. Results of analyses where congruency conditions are combined mirror the results of the trials separated by congruency condition (see Supplementary Table 4).

The second two regression models predict growth in math achievement from nonsymbolic number comparison accuracy rate for Congruent trials (**Table 8**) and Incongruent trials (**Table 9**) by entering math achievement at Time 1 as the first predictor in each model. As noted in the measures, we preregistered an analysis predicting TEMA-3 scores at Time 2 controlling for TEMA-3 at Time 1. However, the sample was missing a substantial number of TEMA-3 scores at Time 1 (n = 16 of 65, or 25%). Therefore, we used the ASPENS Basic Arithmetic Facts and Base 10 (Basic Facts) subtest as a measure of math achievement at Time 1. Model 1 in each regression shows the relation between nonsymbolic number comparison and math achievement at Time 2 controlling for math achievement at Time 1. Model 2, similar to

Table 6. Regression Model Predicting Math Achievement (ASPENS: Basic Facts) at Time 1 from Congruent Nonsymbolic Number Comparison Trials (Model 1) and Nonsymbolic Number Comparison With Additional Measures (Model 2) (n = 65).

Variable	Model 1			Model 2		
	β	95% CI		β	95% CI	
		Low	Up		Low	Up
NNC Congruent (T1)	0.292*	0.052	0.533	0.071	-0.133	0.274
ASPENS: Symbolic MC (T1)				0.598***	0.348	0.849
HTKS				0.162	-0.040	0.363
Oral Reading Fluency				0.004	-0.245	0.253
R^2	0.085			0.467		
ΔR^2				0.381		
F for change in R^2				14.3***		

Note. Regression coefficients are standardized. CI = Confidence Interval; NNC = Nonsymbolic Number Comparison; MC = Magnitude Comparison; HTKS = Head, Toes, Knees, Shoulders.
* p < .05, **p < .01, ***p < .001

Table 7. Regression Model Predicting Math Achievement (ASPENS: Basic Facts) at Time 1 from Incongruent Nonsymbolic Number Comparison Trials (Model 1) and Nonsymbolic Number Comparison With Additional Measures (Model 2) (n = 65).

Variable	Model 1			Model 2		
	β	95% CI		β	95% CI	
		Low	Up		Low	Up
NNC Incongruent (T1)	0.310*	0.071	0.549	0.126	-0.070	0.322
ASPENS: Symbolic MC (T1)				0.588***	0.341	0.836
HTKS				0.159	-0.037	0.355
Oral Reading Fluency				0.002	-0.245	0.248
R^2	0.096			0.477		
ΔR^2				0.381		
F for change in R^2				14.6***		

Note. Regression coefficients are standardized. CI = Confidence Interval; NNC = Nonsymbolic Number Comparison; MC = Magnitude Comparison; HTKS = Head, Toes, Knees, Shoulders.
* p < .05, **p < .01, ***p < .001

Table 8. Regression Model Predicting Growth in Math Achievement (TEMA-3) from Time 1 Nonsymbolic Number Comparison Accuracy Rate on Congruent Trials (Model 1) and Nonsymbolic Number Comparison With Additional Measures (Model 2) (n = 65).

	Model 1			**Model 2**		
Variable	**β**	**95% CI**		**β**	**95% CI**	
		Low	**Up**		**Low**	**Up**
ASPENS: Basic Facts (T1)	0.742***	0.566	0.918	0.557***	0.337	0.777
NNC Congruent (T1)	0.021	-0.155	0.197	-0.029	-0.203	0.146
ASPENS: Symbolic MC. (T1)				0.162	-0.089	0.412
HTKS				0.054	-0.121	0.229
Oral Reading Fluency				0.170	-0.042	0.382
R^2	0.560			0.620		
ΔR^2				0.060		
F for change in R^2				3.10*		

Note. Regression coefficients are standardized. CI = Confidence Interval; NNC = Nonsymbolic Number Comparison; MC = Magnitude Comparison; HTKS = Head, Toes, Knees, Shoulders.
* *p* < .05, ***p* < .01, ****p* < .001

Table 9. Regression Model Predicting Growth in Math Achievement (TEMA-3) from Time 1 Nonsymbolic Number Comparison Accuracy Rate on Incongruent Trials (Model 1) and Nonsymbolic Number Comparison With Additional Measures (Model 2) (n = 65).

	Model 1			**Model 2**		
Variable	**β**	**95% CI**		**β**	**95% CI**	
		Low	**Up**		**Low**	**Up**
ASPENS: Basic Facts (T1)	0.748***	0.570	0.925	0.560***	0.338	0.782
NNC Incongruent (T1)	0.001	-0.176	0.178	-0.027	-0.198	0.144
ASPENS: Symbolic MC (T1)				0.160	-0.090	0.409
HTKS				0.051	-0.121	0.223
Oral Reading Fluency				0.170	-0.042	0.382
R^2	0.560			0.620		
ΔR^2				0.060		
F for change in R^2				3.11*		

Note. Regression coefficients are standardized. CI = Confidence Interval; NNC = Nonsymbolic Number Comparison; MC = Magnitude Comparison; HTKS = Head, Toes, Knees, Shoulders.
* *p* < .05, ***p* < .01, ****p* < .001

the previous two regression models, adds the same list of measures to consider the impact of controlling for other factors. Results indicate that none of the factors considered are significant predictors of growth in math achievement. This is true for Incongruent and Congruent trials considered as the only additional predictors in the model beyond math achievement at Time 1 (Model 1), and when additional cognitive factors are added (Model 2). Results of analyses where congruency conditions are combined mirror the results of the trials separated by congruency condition (see Supplementary Table 5).

Since ASPENS Basic Facts was available at Time 2 as well, we repeated the second two regression models predicting growth in math achievement using ASPENS Basic Facts at Time 1 and Time 2 to check how the current results may be affected by the use of a different math achievement measure. Results from these regression analyses mirror those where the outcome is TEMA-3 (i.e. no significant predictors of growth in math achievement) and are available in **Supplementary Table 2** & **3.**

Lastly, in order to check whether the inclusion of participants with a differing number of trials from the mode of the current sample, we conducted supplementary analyses limited to a subset with 72 trials across Time 1 and Time 2. These supplementary results support the current results presented here in the main manuscript (see **Appendix C**).

4. Discussion

The perception of nonsymbolic numerical magnitudes is a foundational cognitive ability that relates to math achievement (Schneider, Beeres, Coban, Merz, Susan Schmidt, Stricker, & De Smedt, 2017). However, there are many outstanding questions about the cognitive mechanisms that influence its development and what this may reveal about its relation to math skills. The current study investigates these issues by: (1) analyzing first grade children's change in performance in the nonsymbolic number comparison task from the middle to end of first grade split by visual congruency condition; (2) exploring what factors predict change in the task, including executive function and symbolic number skills; and (3) relating nonsymbolic number comparison performance to math achievement and growth in math achievement alongside several other key factors typically related to math growth. We found that accuracy rate increases during first grade at a similar rate for trials with congruent and incongruent visual cues, but that only change in congruent trials was predicted by symbolic number skills. This suggests that during this developmental window, there is a sharpening sense of nonsymbolic numerical magnitude that is influenced by existing symbolic number skills. Additionally, we found that nonsymbolic number comparison performance was associated with concurrent math achievement scores, but not growth in math achievement. When controlling for symbolic number skills, executive function ability, and reading fluency, nonsymbolic number comparison performance was no longer a significant predictor of concurrent math achievement. Instead, symbolic number skill was the unique significant predictor. Together, these findings suggest that the cognitive mechanisms associated with nonsymbolic number processing are undergoing development related to existing symbolic number skills, but appear not to be the principle factors driving math gains during this period.

4.1 Evidence for a sharpening sense of numerical magnitude in 1st grade

In the current study, we reasoned that if children's nonsymbolic number perception increased as the result of a sharpening sense of magnitude, children's scores would increase on congruent and incongruent trials at the same rate. On the other hand, if children's nonsymbolic number perception increased as the result of attending to number and filtering out irrelevant visual cues, then children would improve more rapidly on incongruent trials.

Analysis of cross-sectional data of the number comparison task across age groups has supported both accounts of development in separate studies. In a study comparing six age groups, including 3-, 5-, 7-, 9-, 11-year-olds and adults, Odic (2018) found that there was no effect of age group on the congruency effect (i.e. the difference between congruent and incongruent accuracy rates). Given that there was a significant improvement in performance across age groups, results are interpreted to indicate that nonsymbolic numerical magnitude precision develops over time, and further, that this development is not driven by inhibitory control. In another cross-sectional study comparing three age groups (3-6 year-olds, 8-12 year-olds, and adults), Piazza et al. (2018) report contrasting results. Children in the preschool age group had an average accuracy rate 45.9 points higher for congruent trials than incongruent trials, the 8-12 year-old group actually scored 8.5 points higher for *incongruent* trials (albeit with no significant congruency effect), and adults were 23 points higher for congruent trials. These findings point to a rapid improvement on incongruent trials between 3-6 and 8-12 years of age. Further analyses in the study detail the relative weight of numerical and non-numerical stimulus dimensions for predicting trial-by-trial performance across age groups. Those results indicate that the predictive weight of the numerical dimension increases over time, while weight of the non-numerical dimensions (i.e. item surface area, total surface area, field area/convex hull, and sparsity) decreases over time. Together, their findings support an account of nonsymbolic number perception development driven by increased attention to number and an improvement in filtering out non-numerical visual cues.

Results of the current longitudinal study indicate that children's accuracy increased both on congruent and incongruent trials from the middle to the end of 1st grade, and that there was no significant difference between formats in the rate of increase. This finding suggests that, during the developmental window in the current study, the second half of first grade, increasing nonsymbolic number performance is driven by a sharpening sense of numerical magnitude.

There are, however, differences across these studies that prove problematic for a direct comparison. First, whereas the studies by Odic and Piazza et al. span a wide range of cross-sectional data from 3 years-old to adulthood, the current study examines a much shorter developmental window during 1st grade. Therefore, the current results may apply specifically to cognitive development happening during this age and as a response to first grade curriculum, which focused largely on whole number skills, quantity comparisons, number combinations within 20, and visual representations of numerals. Executive functions, symbolic number development, and other cognitive factors that may influence number skills are also developing rapidly during childhood and may be developing at different rates across individuals and ages. More expansive longitudinal data that tracks the development of other relevant cognitive factors will provide a more complete account of developing numerical precision versus enhanced focusing/filtering over time. Second, there appear to be substantial differences across studies in the degree to which congruency affected participant performance. Wheras Odic did not find a significant congruency effect across age groups, Piazza et al. report a sizeable congruency effect at preschool and with adults, but not in the 8-12 year-old group. In the current study, there was a medium effect of congruency at Time 1 (Cohen's $d = 0.512$) and a medium effect at Time 2 (Cohen's $d = 0.612$). These difference are likely to be driven, at least in part, by differences in stimulus design. While all three studies focus control of visual parameters of dot sets principally on the congruency of surface area, which has been shown to be a dominant visual feature (Clayton et al., 2015), other factors, such as dot size variability, density, and degree of visual congruency are also likely to contribute to the differences in findings (Gilmore et al., 2016a). It is so far unclear what drives the difference in congruency effects across studies, but these factors must also be resolved to provide resolution in the ongoing study of nonsymbolic numerical perception development. Further, the developmental processes of sharpening and filtering numerical perception should be studied over a longer period of time. The current measurement window may have been too short to capture and increase in filtering skills. Developmental trajectories related to filtering may unfold in protracted timeframes, or in response to formal math curriculum that was not the focus of the school settings of the children in the current sample.

4.2 Symbolic number skills predict sharpening sense of numerical magnitude

The second main finding from the current study was that symbolic number skills at T1 predicted change in nonsymbolic accuracy rates for congruent trials. This finding is in agreement with a growing body of literature demonstrating the positive effect of acquiring symbolic, exact number systems, or even the improvement of symbolic skills, on nonsymbolic numerical abilities (for comprehensive reviews, see (Goffin & Ansari, 2019) and (Mussolin et al., 2015). The most dominant perspective of the relation between nonsymbolic and symbolic number processing has been that the nonsymbolic numerical magnitude system is the foundation for symbolic numbers (Piazza, 2010), though recent work in cognitive neuroscience has challenged the details of this model (Wilkey & Ansari, 2020). Since the nonsymbolic system is evolutionarily ancient and not culturally dependent, it is intuitive that symbolic numbers would be mapped onto pre-existing nonsymbolic representations of numerical magnitudes. Increasing acuity of the nonsymbolic system, it follows, would also lead to better symbolic number skills such as fluency with Arabic digits and even arithmetic. While some studies have shown support for this trajectory (i.e. nonsymbolic to symbolic influence; (Libertus et al., 2011; Mazzocco et al., 2011; Nosworthy et al., 2013), a number of recent findings with Kindergarten and 1st grade children suggest a more bi-directional developmental relation (Elliott et al., 2019; Toll et al., 2015) or even a reverse unidirectional relation whereby earlier symbolic skills predict later nonsymbolic skills (Kolkman et al., 2013; Lau et al., 2021; Lyons et al., 2018; Matejko & Ansari, 2016; Mussolin et al., 2014). The current results support a symbolic to nonsymbolic influence. And, as this relation is specific to congruent trials, these results further suggest that improvement in nonsymbolic comparison performance is related to a sharpening sense of magnitude, rather than other cognitive factors, such as attention to number and inhibition, that are believed to more heavily influence performance on incongruent trials.

4.3 Nonsymbolic number relates to math achievement, but does not predict growth

When we explored the relation between number comparison skills, bivariate correlations indicated that nonsymbolic comparison was related to arithmetic concurrently at Time 1 (r = .292 for congruent trials and r = 0.310 for incongruent trials), in line with previous reports. Whereas some previous studies have reported a stronger relation between performance on incongruent trials and math (Bugden & Ansari, 2016; Fuhs & McNeil, 2013; Gilmore et al., 2013; Keller & Libertus, 2015; Wilkey et al., 2018) this pattern does not hold for the current results. A stronger relation between incongruent trials and math has been interpreted as an indication that inhibition in the context of nonsymbolic numerical judgement is a driving factor in the relation between numerical representation and math skills. Given that the current study does not find this bias towards prediction from incongruent trials, this positive correlation between math achievement and *both* congruent *and* incongruent trials indicates the correlation is in some way related to numerical processing itself and not simply inhibition.

The list of previous work demonstrating a stronger relation between math achievement and incongruent trials spans pre-K children to early adolescence, so the difference is unlikely to be related only to sample age. This difference could be due to the above-mentioned differences in visual parameter controls that lead to differing degrees of congruency. In the current study, there was no “anti-correlated” condition, where surface area is actually anti-correlated with numerosity and other incongruent properties, such as dots size, are even further exaggerated. Trials where the surface area is “anti-correlated” or “inverse” to numerosity, rather than simply equated, tend to show the greatest difference from congruent trials in their congruency effect and also predictive relation to math (Fuhs & McNeil, 2013). Future studies should include a wider array of congruency conditions or consider indexing congruency as one predictor in stimulus space among other visual cues that influence behavior, an approach used successfully by DeWind and colleagues (DeWind et al., 2015; DeWind & Brannon, 2016; Starr et al., 2017).

Another possibility is that the relation among trials is an important feature to capture when measuring attention to number. For example, while the congruency effect observed in the current study indicates we are capturing the effect of visual cue congruency, accuracy rate during these trials may not adequately capure the range of individual differences in attention to number. In a recent study, Fuhs et al. (2021) had preschool children complete two tasks, one numerical discrimination and one spatial discrimination. Results showed that flexibly attending to numerical magnitude (and spatial magnitude), indexed as performance on trials where children switched from one dimension to the other, was related to both EF and math skills, but that this performance related to math beyond either EF or performance in the Panamath task. Future work should take into account the relation between trials in addition to performance trends within congruency conditions.

In the current study, when other predictors were added to the models predicting concurrent achievement (i.e. executive function, reading fluency, and symbolic comparison) only symbolic number comparison was a significant predictor. This was true for both the model with congruent trials and incongruent trials as predictors. While previous studies regularly report a stronger relation between symbolic number comparison and math than nonsymbolic comparison (De Smedt et al., 2009; Schneider, Beeres, Coban, Merz, Susan Schmidt, Stricker, & De Smedt, 2017), effect sizes from the current study indicate that the relation was nearly twice as strong for symbolic number comparison (around $r \approx 0.3$ for nonsymbolic and $r \approx 0.6$ for symbolic) and that the relation for nonsymbolic was no longer significant after adding in the symbolic comparison predictor. Still, none of the factors measured in the current study predicted change in math achievement from the middle to end of first grade. Many questions remain about what drives the developmental relation between symbolic and nonsymbolic number skills and their relation to math, and caution is warranted given the limited developmental window of the current study. The role of nonsymbolic skills in supporting the acquisition of whole number knowledge, fluency with basic number combinations, and comfort or familiarity with formal math tasks as students acquire these skills in first grade may differ widely from the role of nonsymbolic skills in supporting math achievement in later years. Further, the delay period in the current study is only 4-5 months. It may be that a longer window of development could show even more gains, or gains in different skills, that capture different developmental trajectories in the periods just before and after 1st grade. As students acquire more fluency, begin to apply number understandings to conceptual tasks, and utilize known number and computation skills to complete increasingly complex math activities, nonsymbolic number skills may provide a different support for mathematical thinking.

4.5 Implications for Schools

Results from this study have practical implications within the context of schools and school based decision making. Currently in kindergarten and first grade settings, symbolic number skills tasks are widely used to screen for students at-risk for mathematics difficulties and to monitor their growth (Witzel & Clarke, 2015). Measures include tasks such as identifying numerals, comparing symbolic magnitudes, and identifying the symbolic number missing from a sequence of numbers (Gersten et al., 2012). The field has called for the investigation into other constructs that might lead to better screening and progress monitoring (Methe et al., 2011). Researchers have suggested working memory, student engagement, executive functioning, and nonsymbolic number skills as constructs that warrant exploration (Gersten et al., 2012; McClelland et al., 2014; Nosworthy, Bugden, Archibald, Evans, & Ansari, 2013, Peng & Kievett, 2020).

The findings from this study demonstrate the continued importance of symbolic number skills and their role in screening and monitoring student growth. However, it is clear that nonsymbolic skills continued to develop during first grade and showed significant relationships with math achievement and symbolic number skills, suggesting they may have potential for increasing the precision of traditional screening batteries for some students. Potential applications of increased precision of monitoring tools would include the placement of students identified as at-risk in interventions of varying intensity or interventions specifically designed to account for the role of non-academic constructs. Emerging evidence suggests the potential promise of these approaches in increasing the effectiveness of early interventions (e.g. Al Otaiba et al., 2014; Fuchs et al., 2014; Peng & Fuchs, 2017). Future research should continue to identify and explore additional constructs to develop and validate assessment batteries to improve educational decision-making.

4.6 Other Limitations and Future Directions

A number of limitations in the current study should be noted to guide interpretation of results and improvements to future investigations. First, in this study we measures EF skills with a single task, the Head, Toes, Knees, and Shoulders task. While this task draws on the multiple components of EF (i.e. inhibition, working memory, and rule shifting), it is employed in one task with a single measure. Future studies should aim to combine multiple measures and tasks to capture a broader array of individual differences in these cognitive abilities specific to each component. Nuance in EF measurement may yield more information regarding how EF affects the development of numerical magnitude processing. Further, EF was only measured at one timepoint. Given that EF is under rapid development during 1st grade in addition to numerical and math skills, we were unable to detect whether change in EF relates to the question of sharpening or focusing nonsymbolic number discrimination. With multiple timepoints of each measure, the developmental relations may be more directly investigated. More than two timepoints would also allow us to capture non-linear change in development. Second, while school-district level data indicated that children were sampled from a socio-economically diverse student body, student-level SES data was not available for the current dataset. Some researchers have suggested that there may be differences in the importance of EF for children of low- versus high-SES backgrounds in nonsymbolic number skills (Fuhs et al., 2016; Fuhs & McNeil, 2013) or for the relation between symbolic and nonsymbolic number skills (Sepúlveda et al., 2020). Third, while the current study elicited robust congruency effects at both timepoints, which was the main effect necessary to investigate the question of focusing versus sharpening, we did not conduct a detailed analysis of the various visual parameters that led to a congruency effect. The current study focused on surface area and dot size to control for visual cue congruency. Active research on the various parameters of nonsymbolic number stimuli indicates that some features affect numerical perception more than others (Clayton et al., 2015; DeWind et al., 2015; Rinsveld et al., 2020; Salti et al., 2017). In particular, surface area and even moreso convex hull (the total area subtended by dot stimuli) incongruent with numerosity have been shown to elicit strong congruency effects (Clayton & Gilmore, 2015; Gebuis & Reynvoet, 2011; Gilmore et al., 2016b). The Panamath task used to generate stimuli in the current study accounts for (in)congruencies in surface area but not convex hull (Guillaume et al., 2020), which means the current study is not a complete account of the various parameters related to congruency and leaves open whether stronger congruency effects elicited by stimuli with incongruent convex hull might lead to different results. It may also follow that attention to these various cues may change over time at different rates, which may further inform models of numerical magnitude perception in regards to sharpening, filtering, and allocation of attention. Lastly, most measures used in the current study are timed (i.e. symbolic and nonsymbolic comparison tasks). Exploring untimed versions of these tasks may elicit more variability in task strategy that would relate to existing individual differences in children's EF skills.

4.6 Conclusions

Together, our results indicate that nonsymbolic number comparison improves in 1st grade at a similar rate for trials with varying levels of visual cue congruency in the current task paradigm, suggesting an overall sharpening of magnitude representations. This change was predicted by existing symbolic number skill level. Alongside previous research, this study suggests that children continue to develop their nonsymbolic representations of number in tandem with executive function skills but in a process that is directly related to symbolic number development.

Acknowledgements: This work is supported by the National Science Foundation (NSF) grants DRL 1660840 and DRL 1748954 to Drs. Ben Clarke, Hank Fien, Fred Sabb, and Lina Shanley and by the U.S. Department of Education (USDE) Institute for Education Sciences (IES) grant R324A160046 to Drs. Ben Clarke, Christian Doabler, and Hank Fiend at the University of Oregon. Eric D. Wilkey is the recipient of a Banting Postdoctoral Fellowship (NSERC) and BrainsCAN Postdoctoral Fellowship at Western University, funded by the Canada First Research Excellence Fund (CFREF).

References

Al Otaiba, S., Connor, C. M., Folsom, J. S.,Wanzek, J., Greulich, L., Schatschneider, C., & Wagner, R. K. (2014). To wait in Tier 1 or intervene immediately: A randomized experiment examining first-grade response to intervention in reading. Exceptional Children, 81, 11–27. doi:10.1177/0014402914532234

Bugden, S., & Ansari, D. (2015). Probing the nature of deficits in the ‘Approximate Number System’ in children with persistent Developmental Dyscalculia. *Developmental Science*, 1–17. https://doi.org/10.1111/desc.12324

Bugden, S., & Ansari, D. (2016). Probing the nature of deficits in the ‘Approximate Number System’ in children with persistent Developmental Dyscalculia. *Developmental Science*, *19*(5), 817–833. https://doi.org/10.1111/desc.12324

Butterworth, B. (2012). *Dyscalculia Screener*. https://doi.org/10.1037/t05204-000

Cameron Ponitz, C. E., McClelland, M. M., Jewkes, A. M., Connor, C. M. D., Farris, C. L., & Morrison, F. J. (2008). Touch your toes! Developing a direct measure of behavioral regulation in early childhood. *Early Childhood Research Quarterly*, *23*(2), 141–158. https://doi.org/10.1016/j.ecresq.2007.01.004

Champely, S. (2020). *pwr: Basic Functions for Power Analysis* (R package version 1.3-0) [Computer software]. https://CRAN.R-project.org/package=pwr

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. https://doi.org/10.1016/j.actpsy.2014.01.016

Clarke, B., Gersten, R. M., Dimino, J., & Rolfhus, E. (2011). Assessing student proficiency of number sense (ASPENS). *Longmont, CO: Cambium Learning Group, Sopris Learning*.

Clayton, S., & Gilmore, C. (2015). Inhibition in dot comparison tasks. *ZDM*, *47*(5), 759–770. https://doi.org/10.1007/s11858-014-0655-2

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. https://doi.org/10.1016/j.actpsy.2015.09.007

De Smedt, B., Verschaffel, L., & Ghesquière, P. (2009). The predictive value of numerical magnitude comparison for individual differences in mathematics achievement. *Journal of Experimental Child Psychology*, *103*(4), 469–479. https://doi.org/10.1016/j.jecp.2009.01.010

Dehaene, S. (1997). *The Number Sense*. Oxford University Press.

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. https://doi.org/10.1016/j.cognition.2015.05.016

DeWind, N. K., & Brannon, E. M. (2016). Significant inter-test reliability across approximate number system assessments. *Frontiers in Psychology*, *7*(MAR), 1–10. https://doi.org/10.3389/fpsyg.2016.00310

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L. S., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K., & Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, *43*(6), 1428–1446. https://doi.org/10.1037/0012-1649.43.6.1428

Elliott, L., Feigenson, L., Halberda, J., & Libertus, M. E. (2019). Bidirectional, Longitudinal Associations Between Math Ability and Approximate Number System Precision in

Childhood. *Journal of Cognition and Development*, *20*(1), 56–74. https://doi.org/10.1080/15248372.2018.1551218

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*(1), 53–72. https://doi.org/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. https://doi.org/10.1016/j.tics.2004.05.002

Fox, J., & Weisberg, S. (2018). *car: A Companion ot Applied Regression*.

Fuhs, M. W., 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. https://doi.org/10.1080/15248372.2015.1105228

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. https://doi.org/10.1111/desc.12013

Fuhs, M. W., Tavassolie, N., Wang, Y., Bartek, V., Sheeks, N. A., & Gunderson, E. A. (2021). Children's Flexible Attention to Numerical and Spatial Magnitudes in Early Childhood. *Journal of Cognition and Development, 22*(1), 22–47. https://doi.org/10.1080/15248372.2020.1844712

Geary, D. C., Bailey, D. H., & Hoard, M. K. (2009). Predicting Mathematical Achievement and Mathematical Learning Disability with a Simple Screening Tool. *Journal of*

Psychoeducational Assessment, *27*(3), 265–279. https://doi.org/10.1177/0734282908330592

Gebuis, T., & Reynvoet, B. (2011). Generating nonsymbolic number stimuli. *Behavior Research Methods*, *43*(4), 981–986. https://doi.org/10.3758/s13428-011-0097-5

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), e67374. https://doi.org/10.1371/journal.pone.0067374

Gilmore, C., Cragg, L., Hogan, G., & Inglis, M. (2016a). Congruency effects in dot comparison tasks: Convex hull is more important than dot area. *Journal of Cognitive Psychology*, *28*(8), 923–931. https://doi.org/10.1080/20445911.2016.1221828

Gilmore, C., Cragg, L., Hogan, G., & Inglis, M. (2016b). Congruency effects in dot comparison tasks: Convex hull is more important than dot area. *Journal of Cognitive Psychology*, *28*(8), 923–931. https://doi.org/10.1080/20445911.2016.1221828

Gilmore, C., Keeble, S., Richardson, S., & Cragg, L. (2015). The role of cognitive inhibition in different components of arithmetic. *Zdm*, 1–12. https://doi.org/10.1007/s11858-014-0659-y

Ginsburg, H., & Baroody, A. (2003). *TEMA-3 Examiners manual*.

Goffin, C., & Ansari, D. (2019). How Are Symbols and Nonsymbolic Numerical Magnitudes Related? Exploring Bidirectional Relationships in Early Numeracy. *Mind, Brain, and Education*, *13*(3), 143–156. https://doi.org/10.1111/mbe.12206

Good, R. H., & Kaminski, R. A. (2002). *Dynamic Indicators of Basic Early Literacy Skills (6th ed.)*.

Guillaume, M., Schiltz, C., & Rinsveld, A. V. (2020). NASCO: A New Method and Program to Generate Dot Arrays for Non-Symbolic Number Comparison Tasks. *Journal of Numerical Cognition, 6*(1), 129–147. https://doi.org/10.5964/jnc.v6i1.231

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), 1457–1465. https://doi.org/10.1037/a0012682

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. https://doi.org/10.1073/pnas.1200196109

Halberda, J., Mazzocco, M. M. M. M. M. M. M. M. M., & Feigenson, L. (2008). Individual differences in nonverbal number acuity correlate with maths achievement. [Supplement]. *Nature*, *455*(7213), 8–11. https://doi.org/10.1038/nature

Hibbard, J. H., Peters, E., Dixon, A., Tusler, M., Peters, E., & Dixon, A. (2007). Consumer Competencies and the Use of Comparative Quality Information: It Isn’t Just about Literacy. *Medical Care Research and Review*, *64*(4), 379–394. https://doi.org/10.1177/1077558707301630

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*(May), 1–11. https://doi.org/10.3389/fpsyg.2015.00685

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. https://doi.org/10.1016/j.learninstruc.2012.12.001

Landerl, K., & Kölle, C. (2009). Typical and atypical development of basic numerical skills in elementary school. *Journal of Experimental Child Psychology*, *103*(4), 546–565. https://doi.org/10.1016/j.jecp.2008.12.006

Lau, N. T. T., Merkley, R., Tremblay, P., Zhang, S., De Jesus, S., & Ansari, D. (2021). Kindergarteners' symbolic number abilities predict nonsymbolic number abilities and math achievement in grade 1. *Developmental Psychology*, *57*(4), 471–488. https://doi.org/10.1037/dev0001158

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. https://doi.org/10.1111/j.1467-7687.2011.01080.x

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), 440–457. http://dx.doi.org/10.1037/dev0000445

Matejko, A. A., & Ansari, D. (2016). Trajectories of symbolic and nonsymbolic magnitude processing in the first year of formal schooling. *PLoS ONE*, *11*(3), 1–15. https://doi.org/10.1371/journal.pone.0149863

Mazzocco, M. M. M., Feigenson, L., & Halberda, J. (2011a). Impaired Acuity of the Approximate Number System Underlies Mathematical Learning Disability (Dyscalculia). *Child Development*, *82*(4), 1224–1237. https://doi.org/10.1111/j.1467-8624.2011.01608.x

Mazzocco, M. M. M., Feigenson, L., & Halberda, J. (2011b). Preschoolers' Precision of the Approximate Number System Predicts Later School Mathematics Performance. *PLoS ONE*, *6*(9), e23749. https://doi.org/10.1371/journal.pone.0023749

Mussolin, C., Nys, J., Content, A., & Leybaert, J. (2014). Symbolic Number Abilities Predict Later Approximate Number System Acuity in Preschool Children. *PLOS ONE*, *9*(3), e91839. https://doi.org/10.1371/journal.pone.0091839

Mussolin, C., Nys, J., Leybaert, J., & Content, A. (2015). How approximate and exact number skills are related to each other across development: A review☆. *Developmental Review*, 1–15. https://doi.org/10.1016/j.dr.2014.11.001

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. https://doi.org/10.1371/journal.pone.0067918

Odic, D. (2018). Children's intuitive sense of number develops independently of their perception of area, density, length, and time. *Developmental Science*, *21*(2), 1–15. https://doi.org/10.1111/desc.12533

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), 1103–1112. https://doi.org/10.1037/a0029472

Park, J., & Brannon, E. M. (2013). Training the Approximate Number System Improves Math Proficiency. *Psychological Science*, *24*(10), 2013–2019. https://doi.org/10.1177/0956797613482944

Park, J., & Brannon, E. M. (2014). Improving arithmetic performance with number sense training: An investigation of underlying mechanism. *Cognition*, *133*(1), 188–200. https://doi.org/10.1016/j.cognition.2014.06.011

Piazza, M. (2010). Neurocognitive start-up tools for symbolic number representations. *Trends in Cognitive Sciences*, *14*(12), 542–551. https://doi.org/10.1016/j.tics.2010.09.008

Piazza, M., De Feo, V., Panzeri, S., & Dehaene, S. (2018). Learning to focus on number. *Cognition*, *181*(July), 35–45. https://doi.org/10.1016/j.cognition.2018.07.011

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. https://doi.org/10.1016/j.cognition.2010.03.012

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. https://doi.org/10.1177/0956797612464057

Ponitz, C. C., McClelland, M. M., Matthews, J. S., & Morrison, F. J. (2009). A Structured Observation of Behavioral Self-Regulation and Its Contribution to Kindergarten Outcomes. *Developmental Psychology*, *45*(3), 605–619. https://doi.org/10.1037/a0015365

Price, G. R., Holloway, I. D., Räsänen, P., Vesterinen, M., & Ansari, D. (2007). Impaired parietal magnitude processing in developmental dyscalculia. *Current Biology*, *17*(24), R1042–R1043. https://doi.org/10.1016/j.cub.2007.10.013

Rinsveld, A. V., Guillaume, M., Kohler, P. J., Schiltz, C., Gevers, W., & Content, A. (2020). The neural signature of numerosity by separating numerical and continuous magnitude extraction in visual cortex with frequency-tagged EEG. *Proceedings of the National Academy of Sciences*, *117*(11), 5726–5732. https://doi.org/10.1073/pnas.1917849117

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Salti, M., Katzin, N., Katzin, D., Leibovich, T., & Henik, A. (2017). One tamed at a time: A new approach for controlling continuous magnitudes in numerical comparison tasks. *Behavior Research Methods*, *49*(3), 1120–1127. https://doi.org/10.3758/s13428-016-0772-7

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. https://doi.org/10.1111/desc.12372

Schneider, M., Beeres, K., Coban, L., Merz, S., Susan Schmidt, S., Stricker, J., De Smedt, B., Schmidt, S. S., Stricker, J., Smedt, B. De, Schneider, M., Beeres, K., Coban, L., Merz, S., Schmidt, S. S., Stricker, J., Smedt, B. De, Susan Schmidt, S., Stricker, J., … Smedt, B. De. (2017). Associations of Non-Symbolic and Symbolic Numerical Magnitude Processing with Mathematical Competence: A Meta-analysis. *Developmental Science*, *20*(3), e12372. https://doi.org/10.1111/desc.12372

Sepúlveda, F., Rodríguez, C., & Peake, C. (2020). Differences and Associations in Symbolic and Non-Symbolic Early Numeracy Competencies of Chilean Kinder Grade Children, considering Socioeconomic Status of Schools. *Early Education and Development*, *31*(1), 137–151. https://doi.org/10.1080/10409289.2019.1609819

Smyth, R. E., & Ansari, D. (2020). Do infants have a sense of numerosity? A p-curve analysis of infant numerosity discrimination studies. *Developmental Science*, *23*(2), 1–11. https://doi.org/10.1111/desc.12897

Starr, A., DeWind, N. K., & Brannon, E. M. (2017). The contributions of numerical acuity and non-numerical stimulus features to the development of the number sense and symbolic math achievement. *Cognition*, *168*, 222–233. https://doi.org/10.1016/j.cognition.2017.07.004

Suárez-Pellicioni, M., & Booth, J. R. (2018). Fluency in symbolic arithmetic refines the approximate number system in parietal cortex. *Human Brain Mapping*, *May*. https://doi.org/10.1002/hbm.24223

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*(August 2016), 187–203. https://doi.org/10.1016/j.tine.2016.11.002

Team, R. C. (2018). *R: A Language and Environment for Statistical Computing* (3.5.1). R Foundation for Statistical Computing.

The jamovi project (Version 0.9). (2019).

Toll, S. W. M., Van Viersen, S., Kroesbergen, E. H., & Van Luit, J. E. H. (2015). The development of (non-)symbolic comparison skills throughout kindergarten and their relations with basic mathematical skills. *Learning and Individual Differences*, *38*, 10–17. https://doi.org/10.1016/j.lindif.2014.12.006

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag.

Wickham, H. (2017). *tidyverse: Easily Install andLoad the "Tidyverse"* (R package version 1.2.1).

Wilkey, E. D., & Ansari, D. (2020). Challenging the neurobiological link between number sense and symbolic numerical abilities. *Annals of the New York Academy of Sciences*, *1464*(1), 76–98. https://doi.org/10.1111/nyas.14225

Wilkey, E. D., Pollack, C., & Price, G. R. (2018). Dyscalculia and Typical Math Achievement Are Associated With Individual Differences in Number-Specific Executive Function. *Child Development*, *00*(0), 1–24. https://doi.org/10.1111/cdev.13194

Wilkey, E. D., & Price, G. R. (2018). Attention to number: The convergence of numerical magnitude processing, attention, and mathematics in the inferior frontal gyrus. *Human Brain Mapping*, 1–16. https://doi.org/10.1002/hbm.24422

SUPPLEMENT

Appendix A. District, school, and classroom data for sample.

Districts and schools. Students from eight elementary schools in three suburban Oregon school districts (Districts A, B, and C) participated in this study. The students who participated in years 1 and 2 of the study attended schools in Districts A and B. Year 3 students attended schools in Districts A and C.

District A consists of 11 schools with a total enrollment of 5,500 students. The number of students in each school ranges from 100 to 1,500. In all, 64% of students identify as White, 8% identify as more than one race, and all other racial groups are represented by less than 2% of students. Twenty-two percent of the students identify as Hispanic or Latino. Of the students in District A, 48% qualify for free or reduced lunch, 19% have a special education qualifying disability, and 12% are English learners.

District B consists of 15 schools with a total enrollment of 10,600 students. The number of students in each school ranges from 120 to 1,400. In District B, 67% of students identify as White, 7% identify as more than one race, and all other racial groups are represented by less than 2% of students. Like District A, 22% percent of the students identify as Hispanic or Latino. In District B, 69% of students qualify for free or reduced lunch, 17% have a qualifying disability, and 12% are English learners.

District C consists of 7 schools with a total enrollment of 2,700 students. The number of students in each school ranges from 50 to 780. In all, 82% of students identify as White with all other racial groups represented by less than 2% of students. Twelve percent of the students identify as Hispanic or Latino. In District C, 69% of the students qualify for free or reduced lunch, 19% have a qualifying disability, and 7% are English learners.

Classrooms. The participants in this study were enrolled in classrooms of 25 different teachers. Students in Cohorts 1, 2, and 3 were in 18, 16, and 7 classrooms, respectively. Four teachers participated all 3 years. In year 2, 12 teachers who participated in year 1 returned and 4 new teachers were added. In year 3, 4 teachers continued participating and 3 new teachers were added.

Appendix B. Further details of nonysymbolic number comparison task and analysis steps

See the following link for a full discussion of controlling the visual parameters of dot arrays: http://panamath.org/wiki/index.php?title=Panamath_Software_Manual#How_Size_Controlling_Works). As specified in the preregistration, all participants had greater than 10 trials per congruency condition at each time point. Mean accuracy rates were used as the score of interest. While Weber fractions are often calculated as a discrimination threshold for this task, the calculation relies on enough trials to fit a nonlinear regression model across the data. Splitting performance calculations by congruency condition would not provide enough trials to fit a reliable model (Wilkey et al., 2018), so mean accuracy rates were used. Mean accuracy rates are also a frequently used performance metric for the task and a growing body of literature suggests that mean accuracy is strongly correlated with, and possibly more reliable than, ratio-dependent metrics such as the Weber fraction (Gilmore, Attridge, & Inglis, 2011; Inglis & Gilmore, 2014), a finding which extends to congruency comparisons (Szűcs, Devine, Soltesz, Nobes, & Gabriel, 2013; Szűcs, Nobes, Devine, Gabriel, & Gebuis, 2013).

Difficulty of the Panamath task is partly controlled by the ratio of the dots being presented. In order to adjust task difficulty in a developmentally appropriate manner (i.e. to capture the magnitude detection threshold for multiple age groups) the ratios of the number comparison trials presented are adjusted by age in years. In the current sample, some children turned 1 year older between Time 1 and Time 2 and received a slightly more difficult set of ratios. Children who did not turn a year older received the same version of Panamath at Time 2. In order to adjust for this slight change in difficulty level, we divided all children's accuracy rates by the average ratio they were presented at the time point they were assessed. This results in scores that more adequately reflect children's improvement in performance when they received a more difficult set of trial ratios at Time 2. At the same time, it does not adjust scores for children who received the same version at Time 1 and Time 2. For example, if Child A was 80% correct on a Panamath assessment at Time 1 and the mean ratio presented was 1.88, their adjusted score would be 42.55 (80/1.88 = 42.55). If they scored 85% correct at Time 2 with the same Panamath version (mean ratio = 1.88), their adjusted score would be 45.21 (85/1.88), a difference of 5 accuracy points, or 2.66 adjusted points. If another child, Child B, performed at the same accuracy rate as Child A for both time points (80% at Time 1 and 85% at Time 2), but received a more difficult set of ratios at Time 2 (Time 1 mean ratio = 1.88, Time 2 mean ratio = 1.84), then the difference in accuracy points again would be 5, but the adjusted score would be 3.65 (85/1.84). Therefore, the difference in adjusted points of 2.66 for Child A versus 3.65 for Child B between timepoints represents the adjustment for task difficulty based on ratio. Lastly, in order to make scores more intuitively related to the original accuracy rates, scores were re-scaled to center the variable's mean at the original score's mean accuracy rate. This was done by multiplying scores by the ratio representing the difference in non-adjusted to adjusted scores (original group mean / ratio-adjusted group mean), which has no impact on the relative position of children's scores nor the subsequent analysis.

Supplementary Table 1. Moderator Analysis for Growth in All Trials of Nonsymbolic Number Comparison

Predictor	**M1**	**M2**	**M3**
NNC All Trials (T1)	0.467* [-0.03 – 0.96]	0.788* [0.14 – 1.44]	0.684* [-0.03 – 1.32]
ASPENS: BF (T1)	0.274 [-0.36 – 0.91]		
NNC All Trials (T1) x ASPENS: BF (T1)	-0.051 [-1.02 – 0.72]		
HTKS		0.211 [-0.47 – 0.89]	
NNC All Trials (T1) x HTKS		-0.518 [-1.54 – 0.50]	
ASPENS: Symbolic MC			0.539 [-0.36 – 1.44]
NNC All Trials (T1) x ASPENS: Symbolic MC			-0.500 [-1.74 – 0.74]
R^2	0.228**	.225**	0.241***

Note. Regression coefficients are standardized. 95% confidence intervals are in brackets. NNC = nonsymbolic number comparison; BF = Basic Arithmetic Facts and Base 10 subtest of ASPENS; HTKS = Head, toes, knees, shoulders. ASPENS: Symbolic MC = Magnitude Comparison subtest of ASPENS

*p < .05, **p < .01, ***p < .001

Supplementary Table 2. Regression Model Predicting Growth in Math Achievement (ASPENS: Basic Facts at Time 2) from Time 1 Nonsymbolic Number Comparison Accuracy Rate on Congruent Trials (Model 1) and Nonsymbolic Number Comparison With Additional Measures (Model 2) (n = 65).

	Model 1			**Model 2**		
Variable	**β**	**95% CI**		**β**	**95% CI**	
		Low	**Up**		**Low**	**Up**
ASPENS: Basic Facts (T1)	0.562***	0.355	0.769	0.399**	0.133	0.665
NNC T1 Congruent (T1)	0.153	-0.055	0.360	0.102	-0.108	0.313
ASPENS: Mag. Comp. (T1)				0.105	-0.198	0.407
HTKS				0.072	-0.140	0.284
Oral Reading Fluency				0.194	-0.062	0.450
R^2	0.390			0.446		
ΔR^2				0.056		
F for change in R^2				1.99		

Note. Regression coefficients are standardized. CI = Confidence Interval; NNC = Nonsymbolic Number Comparison; Mag. Comp = Magnitude Comparison; HTKS = Head, Toes, Knees, Shoulders.

* $p < .05$, **$p < .01$, ***$p < .001$

Supplementary Table 3. Regression Model Predicting Growth in Math Achievement (ASPENS: Basic Facts at Time 2) from Time 1 Nonsymbolic Number Comparison Accuracy Rate on Incongruent Trials (Model 1) and Nonsymbolic Number Comparison With Additional Measures (Model 2) (n = 65).

	Model 1			**Model 2**		
Variable	**β**	**95% CI**		**β**	**95% CI**	
		Low	**Up**		**Low**	**Up**
ASPENS: Basic Facts (T1)	0.569***	0.359	0.779	0.391**	0.123	0.660
NNC T1 Incongruent (T1)	0.123	-0.087	0.333	0.091	-0.115	0.298
ASPENS: Mag. Comp. (T1)				0.113	-0.190	0.415
HTKS				0.084	-0.124	0.292
Oral Reading Fluency				0.195	-0.062	0.451
R^2	0.382			0.444		
ΔR^2				0.062		
F for change in R^2				2.20		

Note. Regression coefficients are standardized. CI = Confidence Interval; NNC = Nonsymbolic Number Comparison; Mag. Comp = Magnitude Comparison; HTKS = Head, Toes, Knees, Shoulders.
* p < .05, **p < .01, ***p < .001