Developmental Science

Developmental Science 17:5 (2014), pp 714–726 DOI: 10.1111/desc.12152

PAPER

Numerical predictors of arithmetic success in grades 1–6

Ian M. Lyons,[1] Gavin R. Price,[2] Anniek Vaessen,[3] Leo Blomert [3] and Daniel Ansari[1]

1. *Numerical Cognition Laboratory, Department of Psychology & Brain and Mind Institute, The University of Western Ontario, Canada*
2. *Department of Psychology & Human Development, Peabody College, Vanderbilt University, USA*
3. *Department of Psychology, Maastricht University, The Netherlands*

Abstract

*Math relies on mastery and integration of a wide range of simpler numerical processes and concepts. Recent work has identified several numerical competencies that predict variation in math ability. We examined the unique relations between eight basic numerical skills and early arithmetic ability in a large sample (*N = *1391) of children across grades 1–6. In grades 1–2, children's ability to judge the relative magnitude of numerical symbols was most predictive of early arithmetic skills. The unique contribution of children's ability to assess ordinality in numerical symbols steadily increased across grades, overtaking all other predictors by grade 6. We found no evidence that children's ability to judge the relative magnitude of approximate, nonsymbolic numbers was uniquely predictive of arithmetic ability at any grade. Overall, symbolic number processing was more predictive of arithmetic ability than nonsymbolic number processing, though the relative importance of symbolic number ability appears to shift from cardinal to ordinal processing.*

Introduction

Math, being a complex skill, relies on mastery and integration of a wide range of simpler numerical facts and concepts – both innate and acquired (Butterworth, 1999; Dehaene, 1997; Geary, 2013). It is thus crucial for both practical and theoretical reasons to identify the basic skills that are most predictive of success in math during the early school years.

Recent research has identified a range of basic numerical competencies (e.g. determining which of two quantities is numerically larger; see also Table 1) that predict individual differences in mathematical skills (Bonny & Lourenco, 2013; Booth & Siegler, 2006, 2008; Castronovo & Göbel, 2012; De Smedt, Verschaffel & Ghesquière, 2009; Desoete, Ceulemans, De Weerdt & Pieters, 2012; Durand, Hulme, Larkin & Snowling, 2005; Fuhs & McNeil, 2013; Geary, Hoard, Nugent & Bailey, 2013; Gilmore, McCarthy & Spelke, 2010; Gunderson, Ramirez, Beilock & Levine, 2012; Halberda, Mazzocco & Feigenson, 2008; Halberda, Ly, Wilmer, Naiman & Germine, 2012; Holloway & Ansari, 2009; Kolkman, Kroesbergen & Leseman, 2013; Jordan, Kaplan, Ramineni & Locuniak, 2009; Jordan, Glutting & Ramineni, 2010; Libertus, Feigenson & Halberda, 2011; Libertus, Odic & Halberda, 2012; Libertus, Feigenson & Halberda, 2013; Lonnemanna, Linkersdörfera, Hasselhorna & Lindberg, 2011; Lourenco, Bonny, Fernandez & Rao, 2012; Lyons & Beilock, 2011; Mazzocco, Feigenson & Halberda, 2011a, 2011b; Mundy & Gilmore, 2009; Piazza, Facoetti, Trussardi, Berteletti, Conte, Lucangeli, Dehaene & Zorzi, 2010; Price, Palmer, Battista & Ansari, 2012; Reeve, Reynolds, Humberstone & Butterworth, 2012; Reigosa-Crespo, Valdés-Sosa, Butterworth, Estévez, Rodríguez, Santos, Torres, Suárez & Lage, 2012; Sasanguie, De Smedt, Defever & Reynvoet, 2012; Sasanguie, Göbel, Moll, Smets & Reynvoet, 2013b; Sasanguie, Defever, Maertens & Reynvoet, 2013a). Discovery of such predictors has often been followed by a flurry of claims and counter-claims regarding their relative importance and primacy.

Such divergence and contradiction in the current literature may be due to the fact that the relative importance of different predictors varies as a function of when in development they are assessed. For example, one skill might be of particular importance when children are

Address for correspondence: Daniel Ansari, Department of Psychology, The University of Western Ontario, Westminster Hall, Room 325, London, ON N6A 3K7, Canada; e-mail: daniel.ansari@uwo.ca

Table 1 *The tasks collected and a brief description of each*

Task	Brief description
Arithmetic	Standardized measure of mental arithmetic ability (addition & subtraction).
Counting	Count objects as quickly and accurately as possible
DotComp	Compare two arrays of dots to determine which contains more dots
DotEst	Give a verbal (symbolic) estimate for the number of dots in a briefly presented array
NumComp	Compare two symbolic numbers to determine which represents the larger quantity
NumLine	Mark on a horizontal line labeled 0–100 where a symbolic number should fall
NumOrd	Determine whether three symbolic numbers are in numerically increasing (left–right) order
ObjMatch	Determine which of two sets of household objects matches the number of objects in a third set (sets varied in object composition; counting was permitted)
VisAud	Determine if a spoken number word and an Arabic numeral match
Ravens	Standardized measure of nonverbal (spatial) reasoning/IQ
Reading	Standardized measure of (Dutch) reading ability
StimResp	Press one of four marked squares as quickly as possible

Note: Task abbreviations in Table 1 are used in the remaining tables and in Figure 1.

first learning addition facts in first and second grade while, in subsequent years, other skills may be more important for consolidating and linking these facts to one another to form a more mature and flexible understanding of arithmetic. The majority of previous studies have focused on relatively limited age ranges or on only one or two basic numerical skills, leading to an incohoate patchwork of results from varying ages, measures and sample sizes. The issue is further complicated by the fact that many of these basic skills tend to be strongly related to one another, and studies have differed on whether they assessed the *unique* contribution of a given skill.

To provide a systematic investigation of the role of basic number skills in children's arithmetic competence we assessed children's performance on eight different basic numerical tasks and three non-numerical control tasks that have been identified as predictive of children's early arithmetic skills (see Table 1 for a brief summary of tasks). We collected data from 1391 children in grades 1–6 (> 200 children in each grade). This approach allowed us to assess the unique contributions of each basic numerical skill within each grade, as well as across all six education levels. Our aim was thus to characterize both the forest and the trees when examining the relative importance – at different developmental time-points – of a range of basic numerical skills for early math education. Further, by focusing on unique (instead of zero-order) contributions to arithmetic performance, we can identify the truly fundamental numerical building blocks of early math, rather than simply highlighting relatively incidental predictors that do not bear any specific relation to early math ability.

Methods

Participants

The data collection protocol was approved by the ethics review board at Maastricht University. Data were collected from 1463 Dutch children in grades 1–6. Chance performance is difficult to interpret, so we removed children who performed at chance on any of the tasks for which chance could be defined (> 49% error rate on any of the binary forced-choice numerical tasks: NumComp, DotComp, NumOrd, VisAud, ObjMatch; > 24% error rate on the four-choice StimResp task). This removed 37 children from the analysis (2.53%). To remove outliers, we checked whether a child's score on a given task was more or less than 4 standard deviations from the mean for that task in that grade. If this was the case on any task, the child was removed from further analysis. This removed 35 additional children from the analysis (2.39%). In sum, 72 children (4.92%) were removed from the dataset. The overall final sample size was $N = 1391$ (722 female); grade 1 $n = 208$ (97 female), grade 2 $n = 201$ (104 female), grade 3 $n = 253$ (133 female), grade 4 $n = 252$ (134 female), grade 5 $n = 241$ (126 female), grade 6 $n = 236$ (128 female).

Procedure

Children came from seven different primary schools in the Netherlands. Parents could withhold consent by returning the appropriate form. Trained project workers administered all cognitive measures to each child separately in a quiet room at school. All data were collected in one session for grades 1–2, 5–6; data were collected in two sessions that were never more than 5 days apart for grades 3–4.

The nonverbal intelligence (Ravens) and mathematical achievement (Tempo Test Automatiseren, TTA) tests were paper-and-pencil tests. All other cognitive measures were computerized. In all tasks, children were instructed to respond as quickly and accurately as possible. Several practice trials (3–6) were given for each of the numerical tasks. No feedback was given for any of the tasks during the main experimental trials.

Reading, Ravens and Arithmetic were each scored as the total number of correctly completed items. See the section on Task Scoring below for scoring details on the remaining tasks.

Tasks

Mental arithmetic (Arithmetic)

The Arithmetic task was operationalized using the standardized TempoTest Automatiseren (TTA) of basic arithmetic ability (De Vos, 2010). Worksheets containing 50 addition and 50 subtraction operations were administered to children in all grades. Children were instructed to mentally calculate as many operations as possible within 2 minutes per worksheet. Arithmetic scores were the total number of correctly answered problems across both worksheets. Reported reliability for this task is high (.92; Janssen, Verhelst, Engelen & Scheltens, 2010).

Numeral ordering (NumOrd)

In the NumOrd task, children saw three numbers presented horizontally as Arabic numerals. Half the time, the three numbers were all in numerically increasing order (left–right). In the remaining trials, numbers were either in decreasing or mixed order. Children were instructed to push a button with their left hand if the numbers were all increasing ('in order') or a button with their right hand if they were not ('not in order'). Stimuli remained on the screen until the child responded. There were 28 one-digit trials and 28 two-digit trials. The distances between numbers were roughly evenly divided across trials into distances of 1–3, where absolute distance was always symmetrical around the median number and distance for a given trial was calculated as (max − min)/2. Note that first graders saw only one-digit trials; results were highly similar whether we excluded two-digit trials in NumOrd scores for children in grades 2–6. Reliability on this task was high, for both one-digit ($\alpha = .938$) and two-digit trials ($\alpha = .960$).

Numeral comparison (NumComp)

In the NumComp task, children saw two numbers presented horizontally as Arabic numerals, and their task was to decide which number represented the larger quantity. Children saw 64 trials, 32 of which were one-digit and 32 of which were two-digits. For both sizes, ratios (R = min/max) fell into one of 4 ranges: $R \leq .5$, $R = .5$, $.5 < R < .7$, $R \geq .7$, with eight trials in each ratio-range at each size (one- versus two-digits). Stimuli remained on the screen until the child responded. Reliability on this task was high: $\alpha = .977$.

Dot comparison (DotComp)

In the DotComp task, children saw two arrays of dots – one on either side of the screen – and their task was to decide which array contained more dots. The quantities and ratios used were the same as those in the NumComp task. Stimuli remained on the screen until the child responded (note that strong relations between performance on this task and math ability have been reported previously when allowing for self-paced responses; e.g. Piazza *et al.*, 2010). Reliability on this task was high: $\alpha = .955$.

Due to geometric constraints, within a given trial all versions of a dot-comparison task will allow for at least some non-numerical parameters (such as area, perimeter, density, etc.) to covary with number. This problem is compounded by the fact that participants switch the parameters they rely upon from trial to trial (Gebuis & Reynvoet, 2012), and that there is a linear relationship between the number of parameters either incongruent or congruent with number and the magnitude of the congruency effect. Therefore, paradigms that rely solely on changing the congruency of parameters with number *between* trials may fail to bias participants away from relying on non-numerical parameters. In the current dataset, overall area and average individual dot-size were always incongruent with number (the array with fewer dots had greater overall area and larger average dot-size; individual dot-sizes varied randomly). In other words, the non-numerical strategy available in our study was the more difficult one because relying on it would force children to essentially focus on non-numerical variables that were incongruent with the task goal. Our paradigm therefore relied on the assumption that children would be more likely to rely on the relevant parameter (numerosity) that was congruent with the task goal (identify the numerically larger array) than one that was incongruent (smaller overall area and/or individual dot-size). We saw this assumption as less problematic than the assumption (demonstrated to be highly questionable by Gebuis & Reynvoet) that participants would not switch between the parameters across trials.

Note also that recent work has shown that performance on dot-comparison trials where overall area and average individual dots size are *incongruent* with number (as was the case in our study) is more predictive of math achievement than congruent trials (Gilmore, Attridge, Clayton, Cragg, Johnson, Marlow, Simms & Inglis, 2013). This may have contributed to the relatively large

zero-order effect we observed between DotComp and Arithmetic (Table 3; note that this relation remained highly significant even after controlling for non-numerical factors, Ravens, Reading, StimResp, Age: $p < .001$); however, it is unclear how it explains the null result observed when controlling for numerical factors (Tables 5–6, Figure 1).

Object matching (ObjMatch)

In the ObjMatch task, children were shown a sample array of common objects (various animals and pieces of fruit) and two test arrays of objects below the sample array. The children's task was to determine whether the left or right test array contained the same number of objects as the sample array. Children saw 45 trials in total. On 15 trials, all objects in all arrays were the same. On 15 trials, each of the three arrays contained different types of objects, but the objects within a given array were all the same. On 15 trials, all arrays contained a mixture of object-types. The number of objects in the arrays ranged from 1 to 6, and the absolute numerical distance between the two test arrays was 1 or 2. Stimuli remained on the screen until the child responded. Reliability on this task was high: $\alpha = .956$.

Counting (Counting)

In the Counting task, children saw between 1 and 9 dots, and their task was to count the number of dots on the screen as quickly and accurately as possible. Children were given five trials for each quantity. Trials were scored as correct only if the child's response was exactly correct. Verbal responses were collected by the experimenter in written fashion. Response times were estimated by having the child press a button as they gave their verbal response. Reliability on this task was high: $\alpha = .946$.

Numberline estimation (NumLine)

In the NumLine task, children were shown a horizontal line marked as 0 on the left end and 100 on the right end. On each trial, they were shown an Arabic numeral (centrally presented above the numberline) in the range 3–96 (the number was presented verbally at the same time through a pair of headphones). The children's task was to click (with a computer mouse) on the numberline where they thought the target number should be placed in terms of the relative quantity it represented. Stimuli remained on the screen until the child responded. Children saw 26 total trials. Reliability on this task was high: $\alpha = .940$.

Dot quantity estimation (DotEst)

In the DotEst task, children were shown a single array of dots presented too quickly (750 msec) for the dots to be counted individually, which was followed by a visual mask. The mask remained on the screen until the child responded. Children's task was to estimate the number of dots in the array by giving a verbal response. These responses were manually recorded by the experimenter. Children completed a total of 84 trials (12 each for quantities 1–4, 7, 11, 16). Note that if only trials with target values 4, 7, 11 and 16 were used, results were highly similar. Reliability on this task was good: $\alpha = .824$.

Visual-audio matching (VisAud)

In the VisAud task, children heard a number word spoken aloud and immediately thereafter saw an Arabic numeral on the screen. Note that while this task does not have major precedent in the numerical cognition literature, audiovisual integration has been shown to be highly important in the acquisition of another symbolically mediated complex skill: reading (Blomert, 2011; Blomert & Froyen, 2010). The children's task was to determine whether the numeral and spoken number word were the same quantity (by pressing one of two buttons – left indicating 'same' and right indicating 'different'). Stimuli remained on the screen until the child responded. Children completed 64 trials, 32 in the one-digit range and 32 in the two-digit range. In non-matching cases, both numbers were within the same decade. The ratio between numbers in non-matching cases ranged between .25 and .89. Reliability on this task was high: $\alpha = .973$.

Nonverbal intelligence (Ravens)

The Ravens task comprised a battery of colored progressive matrices. This is a normed, untimed, visuospatial reasoning test for children (Raven, Court & Raven, 1995). Children saw a colored pattern and were asked to select the missing piece out of six choices. Children completed 36 items; a child's score was the total number of correctly completed items. For the Dutch version of this task, Van Bon (1986) reported reliabilities of .80 or higher.

Reading ability (Reading)

The Reading task was part of the normed Maastricht Dyslexia Differential Diagnosis battery (Blomert & Vaessen, 2009), and comprised three subtasks. Subtasks contained high-frequency words, low-frequency words,

or pseudo-words. In each subtask, participants saw a series of up to five screens (advanced by the experimenter), each with up to 15 items (75 total items per task). The children's task was to read each item aloud as quickly and accurately as possible. An experimenter manually marked the accuracy of each item. A child's score on a subtask was the total number of correctly read items in 30 seconds. Scores on the three subtasks were summed to form a child's final Reading score. Reported test–retest reliability for this task is high (.95; Blomert & Vaessen, 2009).

Basic stimulus–response processing (StimResp)

In the StimResp task, children saw four horizontally arranged boxes on the screen. On each of 20 trials, a fish appeared in one of the four boxes. Children's task was to press the corresponding key on the response box as quickly and accurately as possible. Note that chance performance on this task was 25% errors, as opposed to 50% errors in the various binary response tasks described above. Stimuli remained on the screen until the child responded. Reliability on this task was high: α = .944.

Task scoring

Our aim was to systematically assess the unique predictive power of our tasks as simultaneous predictors in a multiple regression model. To ensure compatibility, we thus used a measure of performance that was common to as many of the tasks as possible. For NumComp, DotComp, NumOrd, VisAud, ObjMatch, Counting, and StimResp tasks, we used a composite of error rates and reaction times (correct trials only). Combining measures provided a more complete picture of overall performance in each task, it reduced the number of models and statistical tests needed for the analysis as a whole, thus reducing the risk of false-positives, and it implicitly controlled for any variation in speed–accuracy trade-offs across tasks. Measures were combined according to the formula: $P = RT(1 + 2ER)$, where a higher value indicates worse performance. Error rates were multiplied by 2 because most tasks were binary forced-choice (ER = .5 indicates chance). In essence, one can interpret this measure as reaction times (msec) after they have been penalized for inaccurate performance. The scale runs between a child's actual average response time (where $P = RT$) for perfectly accurate performance (0% errors) and double that value ($P = 2RT$) for chance performance (50% errors). Note that this method is quite similar to inverse efficiency ($I = RT/Accuracy$), with the exception that accuracy is nonlinearly weighted in the case of inverse efficiency (changes in accuracy rates lead to greater change in I the further accuracy rates get from 1). Further, we did not opt to standardize RT and ER and average them together for the simple reason that the method adopted here preserves developmental trends and differences across tasks in a slightly more transparent way (see, e.g. Table 2). Because the dependent measure (Arithmetic) was scored with a higher value indicating better performance, for ease of interpretation, performance scores (P) were multiplied by -1 before being entered into regression analyses, so that a positive relation meant better performance on a given measure was related to better Arithmetic performance. For tasks where a composite measure was used, mean response times and error rates can be found in the Supporting Information (Table S1).

DotEst and NumLine involved many trials on which exactly correct answers are expected to be quite rare, rendering traditional error rates essentially uninterpretable. We instead used percent absolute errors: $PAE = \frac{|Est - Target|}{Scale}$, where *Est* is the child's estimate, *Target* is the target number, and *Scale* is the scale or range of target numbers. For the NumLine task, *Scale* was 100, and for DotEst it was 16. For a given child, final scores were computed by averaging across all trials, with a higher number indicating worse performance. Because the dependent measure (Arithmetic) was scored with a higher value indicating better performance, DotEst and NumLine scores were multiplied by -1 before being entered into regression analyses, so that a positive relation meant that better performance was related to better Arithmetic performance.

Age

To control for age variation within each grade, we included each child's age (fractional years – expressed in true annual calendar cycles from the child's birthday to the testing date) as a predictor in the model. We mean-centered Age within each grade to eliminate the correlation between Age and Grade while preserving Age variation within each grade.

Reliability

Reliability for numerical tasks was computed across the all grades (N = 1391) using all trials in a given task via Cronbach's α. For tasks where we used a composite measure, reliability was computed over composite scores such that response times were doubled for trials where a child committed an error.

Results

Table 2 shows mean performance levels for each of the measures that were entered into the initial regression model (these are computed before flipping ($-1x$) relevant variables; hence, a higher number indicates *worse* performance, with the exception of Arithmetic, Reading, and Ravens). Note that there was a significant effect of Grade for each variable (all ps < .001), such that performance improved as grade increased. The bottom row of Table 2 shows mean Age for each grade.

Table 3 shows zero-order correlations between each continuous predictor and Arithmetic scores; this is shown both for the overall sample (N = 1391) and for each grade separately. Given the large sample size, many of the p-values are vanishingly small; thus effect-sizes (Cohen's ds) are also given in Table 3. (For correlations between numerical tasks, see Supporting Information, Table S2).

Model selection

We used performance on the eight basic numerical and three non-numerical tasks to predict arithmetic scores. Age (mean-centered within each grade) was included as a control measure. We also assessed whether each of these 12 predictors interacted with Grade (6 levels: 1–6). Interaction terms tested for developmental effects by asking whether the slope of the relation between performance on a given task (or age) and Arithmetic performance varied as a function of Grade, controlling for all other effects. This yielded an initial set of 25 predictors (12 continuous main effects, 1 discrete main effect (Grade) and 12 terms whereby each continuous predictor interacted with Grade).

Our aim was an unbiased estimate of the unique Arithmetic variance accounted for by each predictor, so we began by including all 25 predictors in the initial model and then worked backwards in stepwise fashion to find a more parsimonious model fit. Initial-model adjusted R^2 $(\bar{R})^2$ was high at .8058; to avoid overfitting, we then removed the least significant predictor from the model in step-wise fashion until all predictors were at least significant (p < .05). (See Table 4 for details of the model-reduction process.) Model reduction resulted in the complete removal of dot-comparison, visual-audio matching, stimulus–response processing, and a decrease of only .0009 from the initial $(\bar{R})^2$. Note that the last predictor removed was the main effect of dot-comparison. Due to wide interest in this task generated in the recent literature, we elected to retain this predictor in the final model (final $(\bar{R})^2$ was instead reduced by .0008 from initial model fit). The final model is summarized in Table 5. Eight of the 12 interaction terms were removed from the initial model. The four tasks that did show an interaction with Grade were all number-processing tasks.

To examine developmental trends, we assessed the unique predictive capacity of the 10 continuous predictors retained in the final model (see Table 5) on Arithmetic at each grade (i.e. a separate multiple regression model was run at each grade). Partial correlations (and corresponding Cohen's ds) are shown in Table 5. Unique effect-sizes across grades for the seven numerical tasks in the final model are represented in Figure 1.

Table 2 *Performance means for each task in the model at each grade*

	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6
N	208	201	253	252	241	246
Arithmetic	19.7 (.5)	42.2 (1.0)	53.8 (.8)	62.1 (.9)	69.9 (.9)	76.2 (.9)
NumOrd	4945 (127)	3865 (99)	3013 (60)	2626 (54)	2225 (51)	1902 (44)
NumComp	1738 (29)	1290 (22)	1077 (13)	973 (13)	895 (11)	822 (11)
DotComp	1721 (31)	1425 (22)	1355 (22)	1203 (22)	1074 (19)	1037 (18)
ObjMatch	4979 (93)	3749 (69)	3289 (50)	2899 (44)	2558 (40)	2318 (35)
Counting	3665 (62)	2921 (45)	2595 (33)	2282 (32)	2100 (28)	1944 (27)
VisAud	1898 (28)	1422 (21)	1191 (16)	1031 (17)	922 (13)	833 (12)
DotEst	8.46 (0.20)	6.97 (0.17)	6.16 (0.13)	5.88 (0.12)	5.48 (0.11)	5.01 (0.11)
NumLine	14.20 (0.39)	7.47 (0.20)	5.39 (0.11)	4.90 (0.10)	4.52 (0.09)	4.31 (0.09)
Ravens	25.6 (.3)	28.6 (.3)	29.7 (.2)	30.2 (.2)	31.0 (.2)	31.7 (.2)
Reading	49.9 (1.8)	93.7 (1.7)	110.5 (1.3)	124.3 (1.3)	129.1 (1.4)	137.3 (1.3)
StimResp	970 (12)	846 (10)	741 (8)	676 (9)	624 (7)	552 (6)
Age	7.06 (.03)	8.12 (.04)	9.15 (.03)	10.33 (.03)	11.09 (.04)	12.18 (.04)

Note: Values in parentheses are standard errors. For Arithmetic, Ravens and Reading scores, a higher number indicates higher performance. For DotEst and NumLine, values are percent absolute error, so a lower number indicates more accurate performance. The remaining tasks show combined performance (response times and error rates), where a lower number also indicates better performance (response times and error rates for these tasks can be found in the Supporting Information Table S1). The bottom row of Table 2 gives the mean Age for each grade (in years). Note that in the model, Age was mean-centered within each grade.

Table 3 *Zero-order relations between each of the continuous predictors and Arithmetic scores for all grades combined and for each grade separately*

		All grades	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6
N		1391	208	201	253	252	241	236
NumOrd	*d*	1.955	.328	.986	.914	1.121	1.115	1.320
	r_0	.699	.162	.442	.416	.489	.487	.551
	p	2E-204	2E-02	5E-11	5E-12	2E-16	9E-16	4E-20
NumComp	*d*	2.430	1.257	1.249	.922	.994	1.242	1.126
	r_0	.772	.532	.530	.419	.445	.528	.490
	p	8E-276	1E-16	6E-16	4E-12	1E-13	1E-18	1E-15
DotComp	*d*	1.331	.586	.561	.375	.290	.677	.536
	r_0	.554	.281	.270	.184	.143	.321	.259
	p	1E-112	4E-05	1E-04	3E-03	2E-02	4E-07	6E-05
ObjMatch	*d*	2.286	1.040	1.149	.933	1.126	1.365	1.194
	r_0	.753	.461	.498	.423	.491	.564	.512
	p	2E-254	2E-12	5E-14	2E-12	1E-16	1E-21	3E-17
Counting	*d*	2.138	.843	.795	.754	1.157	1.233	1.297
	r_0	.730	.388	.369	.353	.501	.525	.544
	p	4E-232	7E-09	7E-08	8E-09	2E-17	2E-18	1E-19
DotEst	*d*	1.161	.424	.605	.570	.704	.535	.413
	r_0	.502	.207	.290	.274	.332	.258	.202
	p	1E-89	3E-03	3E-05	1E-05	7E-08	5E-05	2E-03
NumLine	*d*	1.859	.864	.837	.484	.639	.653	.708
	r_0	.681	.397	.386	.235	.304	.311	.334
	p	4E-190	3E-09	2E-08	2E-04	8E-07	9E-07	2E-07
VisAud	*d*	2.281	.858	.500	.706	.780	.852	.899
	r_0	.752	.394	.242	.333	.364	.392	.410
	p	1E-253	4E-09	5E-04	6E-08	3E-09	3E-10	5E-11
Ravens	*d*	1.096	.584	.433	.594	.251	.430	.299
	r_0	.481	.280	.212	.285	.125	.210	.148
	p	2E-81	4E-05	3E-03	4E-06	5E-02	1E-03	2E-02
Reading	*d*	2.230	.696	.914	.458	.681	.622	.762
	r_0	.744	.329	.416	.223	.322	.297	.356
	p	6E-246	1E-06	8E-10	3E-04	2E-07	3E-06	2E-08
StimResp	*d*	1.832	.809	.515	.497	.564	.415	.469
	r_0	.675	.375	.249	.241	.272	.203	.228
	p	6E-186	2E-08	4E-04	1E-04	1E-05	2E-03	4E-04
Age	*d*	−.200	.328	−.142	−.288	−.441	−.494	−.579
	r_0	−.099	.162	−.071	−.142	−.215	−.240	−.278
	p	2E-04	2E-02	3E-01	2E-02	6E-04	2E-04	1E-05

Note: A positive correlation (and effect size) indicates that better performance on that task was related to better Arithmetic performance. Note that Age was mean-centered at each grade. Abbreviations: *d* = effect size, r_0 = zero-order *r*-value, *p* = *p*-value.

Table 4 *The progression of model reduction*

Step	Predictor removed	*p*-value of removed predictor	$(\bar{R})^2$after predictor removed	$\Delta(\bar{R})^2$ from init. model
Init.	–	–	.805786	–
1	VisAud	.641	.805577	−.000209
	Grade × VisAud	.207		
2	Grade × DotComp	.541	.805715	−.000071
3	StimResp	.493	.805724	−.000063
	Grade × StimResp	.430		
3	Grade × Age	.358	.805650	−.000136
4	Grade × Reading	.415	.805649	−.000138
5	Grade × DotEst	.371	.805592	−.000194
6	Grade × Ravens	.249	.805355	−.000431
7	**Grade × NumComp**	**.172**	**.804963**	**−.000824**
8	DotComp	.174	.804840	−.000946

Note: In steps 1 and 2, two predictors were removed: the main effect and the corresponding interaction term. The step producing the final model is shown in bold. Abbreviations: Init.: Initial (Model), $(\bar{R})^2$: adjusted $(\bar{R})^2$.

Dot-comparison

It may be argued that including the DotEst and ObjMatch tasks in the model biased us toward a null effect for the DotComp task because all three tasks tap one's approximate, nonsymbolic representation of numerosity. First, it is worth pointing out that we also included several symbolic number tasks, and yet the NumComp task (and several other symbolic tasks) showed unique predictive variance. This implies that the objection is itself biased. Nevertheless, we re-ran the model selection process, but this time omitting the DotEst and ObjMatch tasks (and their corresponding interaction terms) to see if this improved the unique predictive capacity of the DotComp task. It did not. In this case, the DotComp task was in fact the first predictor removed during model selection (p = .626; Grade × DotComp: p = .447). Removing DotComp and the corresponding interaction term slightly increased overall model fit from $(\bar{R})^2$ = .79933 to $(\bar{R})^2$= .79941. The unique contribution of DotComp was not significant at any grade (all ps > .15). Thus, the failure of the DotComp task – a task routinely used to estimate approximate number acuity – to predict unique variance in arithmetic ability cannot be attributed to the fact that two other tasks (DotEst and ObjMatch) included in the initial model also involve nonsymbolic number processing. On a similar note, results for the DotComp task remained nonsignificant even if we removed Num-

Table 5 *Final model results*

Predictor	F	p
NumOrd	63.53	<.001
NumComp	46.53	<.001
DotComp	1.85	.174
ObjMatch	34.04	<.001
Counting	17.67	<.001
DotEst	12.67	<.001
NumLine	42.04	<.001
Ravens	5.78	.016
Reading	37.65	<.001
Age	14.53	.001
Grade × NumOrd	9.80	<.001
Grade × ObjMatch	4.00	.001
Grade × Counting	2.35	.039
Grade × NumLine	4.73	<.001
Grade	55.27	<.001
Intercept	461.47	<.001

Note: Overall $(\bar{R})^2$ = .805 (R^2 = .810). For grade and all interaction terms, numerator df = 5; for all other predictors, numerator df = 1. Error (denominator) df = 1355.

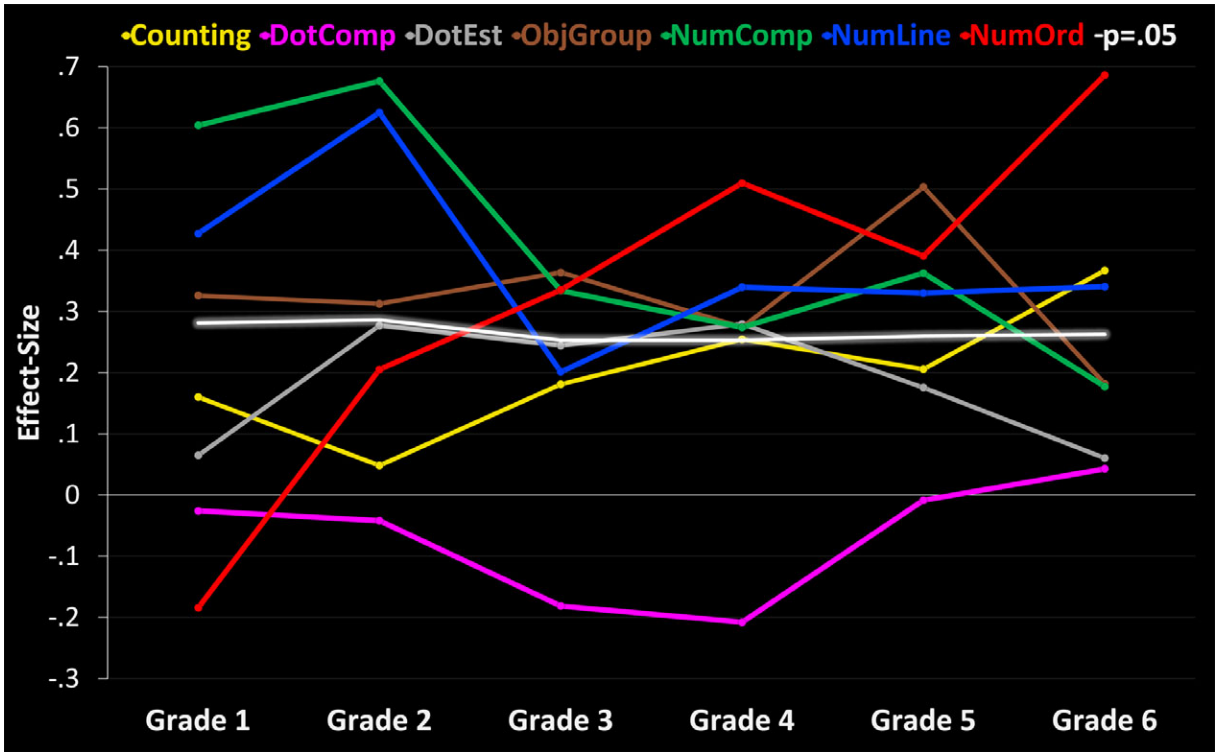

Figure 1 *The change in the relation between the seven numerical tasks from the final model (Table 5) and Arithmetic performance at each grade. Effect sizes correspond to partial-rs taken from the multiple-regression models run at each grade (Table 6), and thus represent unique contributions. The white line is the effect size that corresponds to a partial-r of p = .05 at each grade.*

Comp – which, while symbolic, shares at least the quantity comparison aspect with the DotComp task – from the initial model. In sum, even an approach designed to maximize the potential predictive influence of the DotComp task reveals no evidence that it predicts unique variance in children's arithmetic performance.

Furthermore, for the DotComp task, if we used just error rates (instead of the combined performance measure) results for that task were weaker: the significance of the (already non-significant) partial-*r*s was in fact reduced in each grade. This is important because, typically, only error rates from the DotComp task are used to estimate approximate number acuity (note also that mean error rates are highly correlated with this measure of approximate number acuity; Sasanguie *et al.*, 2012; Szũcs, Nobes, Devine, Gabriel & Gebuis, 2013; Inglis & Gilmore, 2013). In sum, how we calculated performance on the DotComp task is unlikely to explain the lack of any unique relation between DotComp and Arithmetic.

Finally, one might object that some of the DotComp trials use stimuli that fall within the subitizing range (≤ 4), so results may be biased because we are including numerosities whose representation is not truly approximate. Re-running the model using only DotComp trials with numerosities ≥ 10 resulted in complete elimination of the DotComp task from the model in only the 3rd step (as opposed to the 8th step when all DotComp trials were considered; see Table 4). Hence, the inclusion of potentially subitizable stimuli is unlikely to account for the overall lack of unique relation between DotComp and Arithmetic.

Visual-audio matching

Applying logic similar to that for the DotComp task, we tested whether the VisuAud task would be rejected from the final model even if we did not include other tasks that also involved visual presentation of symbolic

Table 6 *Unique relations between each of the continuous predictors from the final model (Table 5) and Arithmetic scores at each grade*

		Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6
N		**208**	**201**	**253**	**252**	**241**	**236**
NumOrd	*d*	−.184	.205	**.335**	**.510**	**.391**	**.686**
	r_p	−.092	.102	**.165**	**.247**	**.192**	**.325**
	p	.198	.159	**.010**	**<.001**	**.003**	**<.001**
NumComp	*d*	**.604**	**.676**	**.334**	**.274**	**.363**	.177
	r_p	**.289**	**.320**	**.165**	**.136**	**.178**	.088
	p	**<.001**	**<.001**	**.010**	**.034**	**.006**	.185
DotComp	*d*	−.026	−.042	−.181	−.208	−.008	.043
	r_p	−.013	−.021	−.090	−.103	−.004	.021
	p	.857	.774	.161	.108	.949	.748
ObjMatch	*d*	**.326**	**.313**	**.364**	**.274**	**.503**	.182
	r_p	**.161**	**.155**	**.179**	**.136**	**.244**	.091
	p	**.023**	**.032**	**.005**	**.034**	**<.001**	.174
Counting	*d*	.160	.048	.181	**.255**	.206	**.367**
	r_p	.080	.024	.090	**.126**	.102	**.181**
	p	.262	.739	.160	**.049**	.120	**.006**
DotEst	*d*	.065	.277	.244	**.280**	.176	.061
	r_p	.033	.137	.121	**.138**	.087	.030
	p	.648	.058	.059	**.031**	.184	.650
NumLine	*d*	**.427**	**.625**	.201	**.340**	**.330**	**.341**
	r_p	**.209**	**.298**	.100	**.167**	**.163**	**.168**
	p	**.003**	**<.001**	.118	**.009**	**.013**	**.011**
Ravens	*d*	.109	.145	**.303**	−.039	.213	.021
	r_p	.054	.073	**.150**	−.020	.106	.011
	p	.446	.318	**.019**	.760	.108	.875
Reading	*d*	**.354**	**.543**	.171	.192	**.311**	**.463**
	r_p	**.174**	**.262**	.085	.096	**.153**	**.226**
	p	**.014**	**<.001**	.186	.137	**.019**	**.001**
Age	*d*	.153	−.096	−.116	−.241	**−.288**	**−.290**
	r_p	.076	−.048	−.058	−.120	**−.143**	**−.144**
	p	.283	.511	.368	.062	**.030**	**.030**

Note: Multiple regression models with all 10 predictors were run at each grade; thus, all statistics represent unique contributions to the model at that grade. A positive correlation (and effect size) indicates that better performance on that task was related to better Arithmetic performance.
Abbreviations: *d* = effectsize, r_p = partial-*r*-value, *p* = *p*-value. Values in bold text are significant at $p < .05$. The top row shows the number of participants (*N*) in each grade.

numbers (NumComp, NumOrd, NumLine tasks). VisAud obtained at least marginal significance (p = .076) in the final model only when NumOrd was omitted from the initial model, indicating that it was rejected from the model in Table 5 due to high collinearity with the NumOrd task (see also Supporting Information Table S2).

Discussion

The present results show that basic symbolic number processing accounts for the majority of unique variance in children's arithmetic ability in grades 1–6, and that the nature of this relationship changes dynamically across grades. The five tasks that showed a significant interaction with grade were all number processing tasks (numeral ordering, object-group matching, number-line estimation, counting, dot-estimation). Figure 1 shows that number-line estimation was a strong unique predictor of arithmetic ability in early grades (1–2), though this fell off from grade 3 on. An examination of Table 6 shows that numeral comparison was also a strong predictor of Arithmetic in grades 1–2. Consistent with prior work (Booth & Siegler, 2006, 2008; Castronovo & Göbel, 2012; De Smedt *et al*., 2009; Desoete *et al*., 2012; Durand *et al*., 2005; Gilmore *et al*., 2010; Gunderson *et al*., 2012; Holloway & Ansari, 2009; Jordan *et al*., 2009, 2010; Kolkman *et al*., 2013; Lonnemanna *et al*., 2011; Mundy & Gilmore, 2009; Reeve *et al*., 2012; Reigosa-Crespo *et al*., 2012; Sasanguie *et al*., 2012, 2013b), our results indicate that assessing the relative magnitude of symbolic numbers – both by directly comparing them with one another and by mapping them onto a visuo-spatial number-line – is perhaps the most important basic numerical ability in the early stages of learning arithmetic skills.

By contrast, numeral ordering, while a poor predictor of arithmetic ability in grade 1, showed a steady increase in unique predictive capacity until grade 6, at which point it was the strongest among all predictors. This is the first demonstration that symbolic number-ordering ability is highly predictive of more complex math skills in children, and is consistent with prior work showing that performance on this task is a strong predictor of complex mental arithmetic ability in adults (Lyons & Beilock, 2011). As the symbolic number system matures, it may be that mastery of basic arithmetic skills increasingly relies on accessing the ordinal information in numerical symbols as opposed to information about relative magnitude (i.e. cardinality). Note that this shift from cardinal to ordinal processing is unlikely to be due to the fact that the NumOrd task was simply more difficult than the NumComp task. First, NumOrd was not the most difficult task at any grade (see Table 2), and it became a stronger predictor even as overall performance was improving. Second, if the NumOrd task were more difficult for reasons unrelated to number processing (e.g. because it tapped into domain-general factors such as working memory or basic processing speed), then controlling for Ravens and StimResp should have eliminated the relation between NumOrd and Arithmetic.

Interestingly, in grades 3–5, it is difficult to identify any one basic numerical skill that supersedes the others. One possibility in keeping with a suggestion made by Kolkman *et al*. (2013) is that basic numerical skills are undergoing a kind of consolidation process at this point. It may be that disparate basic skills are being realigned to support a broader sense of mathematical understanding.

A critical result of the present study is that dot-comparison (DotComp) performance did not predict unique arithmetic variance, either when considering all six grades at once (N = 1391; see Table 5), or in any of the individual grades (Table 6). This is important because many studies have shown that performance on DotComp tasks akin to the one we administered here is predictive of various math abilities (Bonny & Lourenco, 2013; Desoete *et al*., 2012; Gilmore *et al*., 2010; Halberda *et al*., 2008, 2012; Libertus *et al*., 2011, 2012, 2013; Lonnemanna *et al*., 2011; Lourenco *et al*., 2012; Lyons & Beilock, 2011; Mazzocco *et al*., 2011a, 2011b; Piazza *et al*., 2010). None of those studies controlled for the range of other basic numerical *and* non-numerical factors that we do here; further, only one study considered a larger sample (Halberda *et al*., 2012), and that study controlled only for self-reported general IQ. It is worth noting, then, that the zero-order correlations in Table 2 show a significant relation between dot-comparison and mental arithmetic at most grades in our sample. From this, one might well conclude – as many previous researchers have – that dot-comparison ability (and by extension, approximate number processing) is indeed an important precursor for more complex math processing. The partial correlations in Table 6 show that this conclusion is premature. Consistent with prior work (Castronovo & Göbel, 2012; Fuhs & McNeil, 2013; Holloway & Ansari, 2009; Kolkman *et al*., 2013; Sasanguie *et al*., 2013a, 2013b), we find that, once one controls for other basic numerical abilities, this relation is eliminated. Furthermore, even when the other nonsymbolic tasks related to approximate number processing (dot-estimation and object-group matching) were removed from the model, dot-comparison remained a non-significant predictor of arithmetic ability in all respects. Our results therefore provide a strong caution to claims about the importance of approximate number processing for more complex math skills. It is important to

note that our data cannot speak to the role of approximate number processing prior to the onset of formal schooling (though for similar results in children as young as 4, see Fuhs & McNeil, 2013; Kolkman *et al*., 2013; Sasanguie *et al*., 2013a). Regardless, at least by the age of roughly 6–7 years, the importance of the approximate number system is largely overshadowed by other basic numerical and cognitive abilities. On a cognitive level, this implies that the underlying processes or representations involved in these other numerical tasks are more uniquely and directly bound to those that underlie arithmetic (in children grades 1–6). On a practical level, if one could choose only one or two tasks, say, for a diagnostic battery in an educational setting, the zero-order correlations (Table 3) indicate that the DotComp task would not be a bad choice; but based on the multiple regression results (e.g. Figure 1), other numerical tasks might prove more effective – though exactly which task one might choose would depend on the grade in question.

Our results demonstrate for the first time the dynamic relationship between basic numerical abilities and early arithmetic skills over the early elementary grades. This underscores the importance of considering developmental changes when drawing conclusions about the basic numerical foundations of more complex mathematics. This work may help economize and focus future longitudinal studies on specific numerical skills and developmental time-points to allow for stricter causal inferences about the relation between specific basic numerical and math abilities at different points in development. As such, this work offers a unifying perspective through which to interpret the current landscape of disjointed and sometimes contradictory results relating to the cognitive foundations of arithmetic. This work may also help better inform efforts to bridge educational practices with developmental cognitive research.

Acknowledgements

This research was supported by funding from the Canadian Institutes of Health Research (CIHR), The National Sciences and Engineering Research Council of Canada (NSERC), and Canada Research Chairs program (CRC) to Daniel Ansari. Data collection costs were paid in part by Boom Test Uitgevers Amsterdam BV.

References

Blomert, L. (2011). The neural signature of orthographic-phonological binding in successful and failing reading development. *NeuroImage*, **57** (3), 695–703.

Blomert, L., & Froyen, D. (2010). Multi-sensory learning and learning to read. *International Journal of Psychophysiology*, **77** (3), 195–204.

Blomert, L., & Vaessen, A. (2009). *Differentiaal Diagnostiek van Dyslexie: Cognitieve analyse van lezen en spellen [Dyslexia differential diagnosis: Cognitive analysis of reading and spelling]*. Amsterdam: Boom Test Publishers.

Bonny, J.W., & Lourenco, S.F. (2013). The approximate number system and its relation to early math achievement: evidence from the preschool years. *Journal of Experimental Child Psychology*, **114** (3), 375–88.

Booth, J.L., & Siegler, R.S. (2006). Developmental and individual differences in pure numerical estimation. *Developmental Psychology*, **42** (1), 189–201.

Booth, J.L., & Siegler, R.S. (2008). Numerical magnitude representations influence arithmetic learning. *Child Development*, **79** (4), 1016–1031.

Butterworth, B. (1999). *What counts: How every brain is hardwired for math*. New York: The Free Press.

Castronovo, J., & Göbel, S.M. (2012). Impact of high mathematics education on the number sense. *PLoS One*, **7** (4), e33832.

Dehaene, S. (1997). *The number sense: How the mind creates mathematics*. New York: Oxford University Press.

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.

Desoete, A., Ceulemans, A., De Weerdt, F., & Pieters, S. (2012). Can we predict mathematical learning disabilities from symbolic and non-symbolic comparison tasks in kindergarten? Findings from a longitudinal study. *British Journal of Educational Psychology*, **82** (1), 64–81.

De Vos, T. (2010). *Tempo Test Automatiseren*. Amsterdam: Boom Test Publishers.

Durand, M., Hulme, C., Larkin, R., & Snowling, M. (2005). The cognitive foundations of reading and arithmetic skills in 7- to 10-year-olds. *Journal of Experimental Child Psychology*, **91** (2), 113–136.

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.

Geary, D.C. (2013). Early foundations for mathematics learning and their relations to learning disabilities. *Current Directions in Psychological Science*, **22** (1), 23–27.

Geary, D.C., Hoard, M.K., Nugent, L., & Bailey, D.H. (2013). Adolescents' functional numeracy is predicted by their school entry number system knowledge. *PLoS One*, **8** (1), e54651.

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

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.

Gilmore, C.K., McCarthy, S.E., & Spelke, E.S. (2010). Non-symbolic arithmetic abilities and mathematics achievement in the first year of formal schooling. *Cognition*, **115** (3), 394–406.

Gunderson, E.A., Ramirez, G., Beilock, S.L., & Levine, S.C. (2012). The relation between spatial skill and early number knowledge: the role of the linear number line. *Developmental Psychology*, **48** (5), 1229–1241.

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, USA*, **109** (28), 11116–11120.

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

Holloway, I.D., & Ansari, D. (2009). Mapping numerical magnitudes onto symbols: the numerical distance effect and individual differences in children's mathematics achievement. *Journal of Experimental Child Psychology*, **103** (1), 17–29.

Inglis, M., & Gilmore, C. (2013). Sampling from the mental number line: how are approximate number system representations formed? *Cognition*, **129** (1), 63–69.

Janssen, J., Verhelst, N., Engelen, R., & Scheltens, F. (2010). *Wetenschappelijke verantwoording van de toetsen LOVS rekenen-wiskunde voor groep 3 tot en met 8 [Scientific justification of the mathematical test for grades 1 to 6]. Arnhem*, The Netherlands: Cito.

Jordan, N.C., Glutting, J., & Ramineni, C. (2010). The importance of number sense to mathematics achievement in first and third grades. *Learning and Individual Differences*, **20** (2), 82–88.

Jordan, N.C., Kaplan, D., Ramineni, C., & Locuniak, M.N. (2009). Early math matters: kindergarten number competence and later mathematics outcomes. *Developmental Psychology*, **45** (3), 850–867.

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.

Libertus, M.E., Feigenson, L., & Halberda, J. (2011). Preschool acuity of the approximate number system correlates with school math ability. *Developmental Science*, **14** (6), 1292–1300.

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

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

Lonnemanna, J., Linkersdörfera, J., Hasselhorna, M., & Lindberg, S. (2011). Symbolic and non-symbolic distance effects in children and their connection with arithmetic skills. *Journal of Neurolinguistics*, **24** (5), 582–591.

Lourenco, S.F., Bonny, J.W., Fernandez, E.P., & Rao, S. (2012). Nonsymbolic number and cumulative area representations contribute shared and unique variance to symbolic math competence. *Proceedings of the National Academy of Sciences, USA*, **109** (46), 18737–18742.

Lyons, I.M., & Beilock, S.L. (2011). Numerical ordering ability mediates the relation between number-sense and arithmetic competence. *Cognition*, **121** (2), 256–261.

Mazzocco, 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.

Mazzocco, M.M., Feigenson, L., & Halberda, J. (2011b). Preschoolers' precision of the approximate number system predicts later school mathematics performance. *PLoS One*, **6** (9), e23749.

Mundy, E., & Gilmore, C.K. (2009). Children's mapping between symbolic and nonsymbolic representations of number. *Journal of Experimental Child Psychology*, **103** (4), 490–502.

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.

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

Raven, J., Court, J.H., & Raven, J.C. (1995). *Coloured progressive matrices*. Oxford: Oxford Psychologists Press.

Reeve, R., Reynolds, F., Humberstone, J., & Butterworth, B. (2012). Stability and change in markers of core numerical competencies. *Journal of Experimental Psychology: General*, **141** (4), 649–666.

Reigosa-Crespo, V., Valdés-Sosa, M., Butterworth, B., Estévez, N., Rodríguez, M., Santos, E., Torres, P., Suárez, R., & Lage, A. (2012). Basic numerical capacities and prevalence of developmental dyscalculia: the Havana Survey. *Developmental Psychology*, **48** (1), 123–135.

Sasanguie, D., De Smedt, B., Defever, E., & Reynvoet, B. (2012). Association between basic numerical abilities and mathematics achievement. *British Journal of Developmental Psychology*, **30** (Pt 2), 344–357.

Sasanguie, D., Defever, E., Maertens, B., & Reynvoet, B. (2013a). The approximate number system is not predictive for symbolic number processing in kindergarteners. *Quarterly Journal of Experimental Psychology* [epub ahead of print].

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

Szűcs, D., Nobes, A., Devine, A., Gabriel, F.C., & Gebuis, T. (2013). Visual stimulus parameters seriously compromise the

measurement of approximate number system acuity and comparative effects between adults and children. *Frontiers in Psychology*, **4**, 444.

Van Bon, W.H.J. (1986). *Raven's Colored ProgressiveMatrices: Nederlandse normen en enige andere uitkomsten van onderzoek [Raven's Colored Progressive Matrices: Dutch norms and other results] Lisse*. Netherlands: Swets Test Services.

Received: 27 May 2013
Accepted: 7 October 2013

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. Raw means at each Grade (ER: error-rates; RT: reaction-times) on tasks for which a composite performance measure was computed. Values in parentheses are standard-errors.

Table S2. Correlations (and their corresponding effect-sizes, *d*) between numerical tasks. Values below the main diagonal indicate zero-order correlations. Values above the main diagonal indicate partial-correlations controlling for non-numerical factors [Reading, Ravens, StimResp, and Age; df=1385]. Note that Age was non-centered, so the partial-correlations are independent of concomitant improvements in performance across tasks expected in older children (see, e.g., Table S1). All correlations were significant at p<.001, with the exception of the partial-correlation between DotComp and DotEst (p=.673).