

Running head: MEASURING THE APPROXIMATE NUMBER SYSTEM

Measuring the Approximate Number System

Camilla Gilmore

Learning Sciences Research Institute, University of Nottingham

Nina Attridge & Matthew Inglis

Mathematics Education Centre, Loughborough University

Camilla Gilmore

Learning Sciences Research Institute

University of Nottingham

Jubilee Campus Wollaton Road

Nottingham NG8 1BB

UK

camilla.gilmore@nottingham.ac.uk

Abstract

Recent theories in numerical cognition proposed the existence of an approximate number system (ANS) that supports the representation and processing of quantity information without symbols. It has been claimed that this system is present in infants, children and adults, supports learning of symbolic mathematics and that correctly harnessing the system during tuition will lead to educational benefits. Various experimental tasks have been used to investigate individuals' ANSs and it has been assumed that these tasks measure the same system. We tested the relationship across six measures of the ANS. Surprisingly, despite typical performance on each task, adult participants' performances across the tasks were not correlated and estimates of the acuity of individuals' ANSs from different tasks were unrelated. These results highlight methodological issues with tasks typically used to measure the ANS and call into question claims that individuals use a single system to complete all these tasks.

Measuring the approximate number system

How do we represent and process numerical information? Recent influential theories in numerical cognition have proposed the existence of a non-symbolic number system that supports the representation and processing of quantity information without symbols (Cordes, Gelman, Gallistel & Whalen, 2001; Dehaene 1997; Feigenson, Dehaene & Spelke, 2004). The approximate number system (or ANS) is a universal system shared by infants (Xu & Spelke, 2000), children (Barth, La Mont, Lipton & Spelke, 2005), adults from diverse cultures (Pica, Lemer, Izard & Dehaene, 2004), non-human primates (Flombaum, Junge & Hauser, 2005) and other animals (Meck & Church, 1983). The ANS allows individuals to compare, add, subtract and order sets of items (e.g. objects, dots, tones, actions) and the acuity of children's ANS has been related to their performance of formal mathematics (Halberda, Mazzocco & Feigenson, 2008). It has been suggested that educational approaches harnessing the ANS may be effective in supporting children's learning of mathematics (Barth et al., 2005; Gilmore, McCarthy & Spelke, 2007).

Evidence for the existence of the ANS comes from performance on a range of non-symbolic and symbolic tasks. Adults and children can compare, add and subtract sets of dots or sequences of tones with above-chance accuracy. Representations of quantity in the ANS are approximate and the precision of a representation is related to the size of the set – smaller quantities are represented more precisely than larger quantities (see Dehaene, Izard, Spelke & Pica, 2008 and Cantlon, Cordes, Libertus & Brannon, 2009 for discussion of logarithmic vs. linear models of this effect). A characteristic of performance on non-symbolic tasks measuring the ANS is therefore a ratio (or distance) effect on accuracy or response time: performance improves if there is a greater ratio (or numerical distance) between the items to be compared (e.g. Barth, Kanwisher & Spelke, 2003, Cordes et al., 2001; Pica et al., 2004). Similar ratio effects are also observed when adults or children are asked to compare, add or subtract symbolic numerical representations (e.g. digits; Gilmore, et al., 2007; Holloway & Ansari, 2008;

Moyer & Landauer, 1967; Temple & Posner, 1998). This is thought by many researchers to reflect the influence of the ANS through a mapping between non-symbolic and symbolic representations. When we learn exact, symbolic numerical representations (e.g. 5, “five”) these become mapped onto preexisting ANS representations and thus the ANS continues to play a role when individuals use or manipulate symbolic numerical representations (Dehaene & Akhavein, 1995; Moyer & Landauer, 1967; Temple & Posner, 1998). However, other interpretations of the ratio effect exist (Verguts & Fias, 2004; Zorzi & Butterworth, 1999).

Aside from the characteristic ratio signature, there is little previous behavioral evidence that performance across different tasks used to investigate the ANS reflects the existence of a single system. At a minimum, the proposal of a single system used to solve a range of tasks would predict that individual differences in performance should correlate across the tasks. Previous studies have used different tasks to examine the ANS, including small-numerosity comparison tasks with varying numerical distances (e.g. symbolic: Moyer & Landauer, 1967; non-symbolic: Holloway & Ansari, 2008), large-numerosity comparison tasks with varying ratios (e.g. symbolic: Dehaene, Dupoux & Mehler, 1990; non-symbolic: Barth et al., 2003), and approximate addition tasks (e.g. symbolic: Gilmore, et al., 2007; non-symbolic: Barth et al., 2005). Few studies have made use of multiple measures of the ANS and correlations in performance across tasks are frequently not reported. The evidence that does exist provides mixed support for the prediction that performance across tasks will be correlated. Gilmore, McCarthy & Spelke (in review) found that 5- to 6-year-old children’s accuracy on non-symbolic and symbolic versions of an approximate addition task was correlated. Iuculano, Tang, Hall & Butterworth (2008) reported that 8- to 9-year-old children’s performance on approximate non-symbolic addition and comparison tasks was correlated when performance was measured with an efficiency measure (incorporating accuracy and RT), but that there was no such relationship when accuracy alone was considered (T.

Iuculano, personal communication, 28th June 2010). Finally, when performance was measured using the numerical distance effect (NDE; an index of how much performance is reduced when comparing pairs with a small rather than large numerical distance), no correlations were found between NDEs elicited by symbolic and non-symbolic versions of a small-numerosity comparison task in either adults (Maloney, Risko, Preston, Ansari & Fugelsang, 2010) or 6- to 8-year-old children (Holloway & Ansari, 2009). Taken together these studies demonstrate that there is currently little evidence that individual differences in performance across a range of tasks designed to examine the ANS are related.

In this study we test this important prediction of ANS theories by measuring adult participants' ANS performance on a range of non-symbolic and symbolic tasks. These tasks are drawn from those typically used in the literature to investigate ANS acuity. We examine the correlations in performance across the set of tasks to test the theory that a single system underlies performance on each task.

Methods

Participants

One hundred and one participants (50 male, aged 18-48, $M=23$) were paid £20 for taking part. Participants were recruited from the Learning Sciences Research Institute participant pool at the University of Nottingham.

Materials and procedure

Participants were tested individually either in a single session of 2 to 3 hours or two shorter testing sessions. Tasks for which the primary measure was accuracy rather than reaction time had response time limits to prevent ceiling effects. Participants were shown a 'Please speed up' message appearing after the time limit had expired, followed by the next trial. Time limits were determined by taking the Mean RT +1SD obtained in pilot testing. All non-symbolic tasks were given first, administered in random order, followed by the symbolic tasks in random order, and finally the intelligence test.

Participants were offered breaks at set intervals in each computer task (usually every 40 trials) and could also take breaks between tasks. Responses were recorded via the leftmost and rightmost buttons on a five button response box.

Non-symbolic large-numerosity comparison. This task measured participants' ability to choose the more numerous of two dot arrays and was based on the task used by Barth et al. (2003). The two dot arrays, one blue and one red, were presented side by side and the color and side of the screen of the correct answer was counterbalanced. The numerosities of the dot arrays were spread evenly through the range 9-70 and on each trial the arrays differed by one of three ratios (5:6, 5:7, 5:8). Each of the 120 experimental trials began with a fixation point for 1000ms, followed by the dot arrays. Participants were asked to select, as quickly as possible, which array was more numerous. The response time limit was 1249ms.

To prevent participants using strategies based on continuous quantities correlated with number (dot size, luminance, total enclosure area), the stimuli were created following the method of Pica et al. (2004). For each problem two sets of stimuli were created: one in which the dot size and total enclosure area decreased with numerosity, and one in which the dot size and total enclosure area increased with numerosity.

Non-symbolic addition. This task measured participants' ability to compare the sum of two dot arrays to a comparison array and was based on the task used by Barth et al. (2003). The stimuli consisted of three dot arrays presented sequentially - two addend arrays and one comparison array. The order and colors of the addend and comparison arrays were counterbalanced by participant. The problems were created from the pairs of arrays used in the non-symbolic large-numerosity comparison task by creating two new addend arrays, the sum of which matched the size of one of the arrays. The addends were randomly chosen in such a way that neither was larger than the comparison array, or less than 7.

Each of the 120 experimental trials began with a fixation point displayed for 200ms, followed by the first dot array for 500ms, another fixation point for 200ms, the second dot array for 500ms, another fixation point for 200msec, the third dot array for 500ms, and finally a question-mark. Participants were asked to select, as quickly and accurately as possible, whether the sum of the two addend arrays was more or less than the comparison array. The response time limit was 980ms.

Non-symbolic small-numerosity comparison. In this task participants were asked to select the more numerous of two non-symbolic arrays in the number range 1 - 9. This task used the method and stimuli of the non-symbolic comparison task developed by Holloway & Ansari (2009). On each trial two arrays of between 1 and 9 squares were presented simultaneously on the screen. The numerical distance between the arrays ranged from 1 to 6, with equal numbers of trials at each numerical distance. On equal numbers of trials the less numerous array had an overall area which was smaller than, equal to, or larger than the more numerous array and there were also equal numbers of trials in which the less numerous array had either a smaller or larger density than the more numerous array. This prevented participants from reliably using non-numerical aspects of the arrays to base their decisions.

Each of the 72 experimental trials began with a fixation point for 1000ms and then the two arrays were presented side by side and remained until the participant gave a response. Following Holloway & Ansari (2009), a non-symbolic NDE was calculated for each participant on the basis of RT and accuracy

$$\left(\frac{\text{small numerical distance trials} - \text{large numerical distance trials}}{\text{large numerical distance trials}} \right).$$

Symbolic large-numerosity comparison. In this task participants were asked to select the larger of two digits presented simultaneously. The problems were based on those used in the non-symbolic large-numerosity comparison task, but the numerosities were presented as Arabic numerals instead of dots. The pairs of numbers were again

related by the ratios 5:6, 5:7, 5:8, however the numbers now ranged from 10-69 to avoid trials with a one-digit number being compared to a two-digit number. Each of the 120 experimental trials began with a blank screen for 1000ms, followed by the stimuli until response.

Symbolic addition. In this task participants were asked to select the larger of a sum and a comparison number. The problems used were identical to the non-symbolic addition task, but the numerosities were presented as Arabic numerals instead of dots. Participants saw a two-addend sum and a comparison number simultaneously. Each problem appeared four times: $a+b$ vs. c ; $b+a$ vs. c ; c vs. $a+b$; and c vs. $b+a$ with a total of 120 trials. The task was preceded by a practice block of 10 trials. Each trial began with a blank screen for 1000ms, followed by the stimuli. The response time limit was 1279ms.

Symbolic small-numerosity comparison. This task measured participants' ability to select the larger of two digits, in the range 1 - 9. The trials were identical to the non-symbolic small-numerosity comparison task, except that numerosities were presented as Arabic numerals instead of arrays of squares. This task used the method and stimuli of the symbolic comparison task developed by Holloway & Ansari (2009). For each participant a symbolic NDE was again calculated on the basis of RT and accuracy.

Non-verbal intelligence. Participants completed the matrix reasoning subtest of the Wechsler Abbreviated Scale of Intelligence (Wechsler, 1999).

Results

Inclusion criteria

To accurately examine the relationship across measures of the ANS it is important that participants' data is only included if this reflects a true measure of their ANS acuity and not factors such as guessing, or a reliance on non-numerical features of the non-symbolic arrays (e.g. dot size, density etc.). Inclusion of this noisy data might reduce the strength of any relationships found. Therefore, participants' responses on the ANS tasks were examined to remove participants who appeared to be guessing (not

significantly above 50% accuracy) or who appeared to be using non-numerical strategies (a difference in accuracy of $> 50\%$ between the two stimuli sets created by Pica et al's method). A final data set of 61 participants was included in the analysis (10 excluded for guessing, 30 excluded for reliance on non-numerical strategies).¹

Task performance

We used a variety of measures to characterize performance on each task, including accuracy, internal Weber fraction (w), and numerical distance effect (NDE). Our rationale was to use similar measures to those typically used in the literature for each task.

Participants performed above chance on both the non-symbolic large-numerosity comparison ($M=72.7\%$; $t(60) = 23.76$, $p < .001$) and addition tasks ($M=69.1\%$; $t(60) = 15.20$, $p < .001$) and accuracy scores revealed the characteristic linear effect of ratio (comparison $F(1,60) = 218.82$, $p < .001$; addition $F(1,60) = 98.5$, $p < .001$; Figure 1a).

Participants' performance was high and above chance on the symbolic addition task ($M=77.9\%$; $t(60) = 20.32$, $p < .001$) and close to ceiling on the symbolic comparison task ($M=95.6\%$). Nevertheless, accuracy scores on both tasks again revealed a linear effect of ratio (comparison $F(1, 60) = 24.0$, $p < .001$; addition $F(1, 60) = 106.5$, $p < .001$; Figure 1a). Given the high levels of accuracy on the symbolic comparison task, performance was also captured by a measure that integrated both accuracy and response times. For each participant the sum of $1/RT$ of correct trials was calculated. This measure also revealed a linear effect of ratio ($F(1, 60) = 139.8$, $p < .001$).

An alternative method of analyzing accuracy data on approximate tasks is to estimate participants' internal Weber fractions. Participants' accuracy data from the non-symbolic comparison task and the non-symbolic and symbolic addition tasks were individually fitted to the model proposed by Barth, La Mont, Lipton, Dehaene, Kanwisher, & Spelke (2006) using the log-likelihood method. Estimates of the w

parameter for the non-symbolic comparison task ranged from .22 to 1.5 ($M=.43$), for the non-symbolic addition task ranged from .12 to 4.0 ($M=.78$), and for the symbolic addition task ranged from .15 to 1.09 ($M=.38$). Accuracy data from the symbolic comparison task was not modeled because performance was close to ceiling.

Participants' performance on the non-symbolic and symbolic versions of the small-numerosity comparison tasks showed the expected effects of numerical distance on both accuracy (non-symbolic $F(1,60) = 61.57, p < .001$; symbolic $F(1,60) = 56.01, p < .001$; Figure 1) and response times (non-symbolic $F(1,60) = 61.87, p < .001$; symbolic $F(1,60) = 153.41, p < .001$; Figure 1). Non-symbolic NDEs for accuracy ranged from -0.42 to 0.05 ($M=-0.10$) and for RT ranged from -0.09 to 2.23 ($M=0.69$). Symbolic NDEs for accuracy ranged from -0.21 to 0.05 ($M=-0.05$) and for RT ranged from -0.03 to .38 ($M=.15$).

Relationships among tasks

Prior to examining correlations among task performance, the split-half reliability of each task was assessed, using the Spearman-Brown formula. The reliability of each task was found to be acceptable or good (non-symbolic large-numerosity comparison = .846; non-symbolic addition = .929; non-symbolic small-numerosity comparison = .959; symbolic large-numerosity comparison = .782; symbolic addition = .902; symbolic small-numerosity comparison = .958).

The relationships among participants' performance on the 3 non-symbolic tasks was examined using partial correlations, controlling for intelligence (see Table 1).² There was no significant correlation between performance on the non-symbolic large-numerosity comparison task and the non-symbolic addition task when either accuracy or estimates of w were used as the measure of performance (Figure 2a). Perhaps unsurprisingly, participants' NDE for accuracy on the non-symbolic small-numerosity comparison task was correlated with accuracy on the non-symbolic large-numerosity comparison task (Figure 2b); but it was not correlated with accuracy on the non-

symbolic addition task (Figure 2c). Participants' NDE for RT on the non-symbolic small-numerosity comparison task was not correlated with performance on either the non-symbolic large-numerosity comparison task or the non-symbolic addition task.

The relationships among performance on the non-symbolic and the symbolic tasks was then examined using partial correlations, again controlling for intelligence (see Table 2). There was no relationship between performance on the non-symbolic and symbolic large-numerosity comparison tasks (Figure 3a) or between the non-symbolic and symbolic addition tasks (Figure 3b). Finally, participants' NDE for accuracy on the non-symbolic and symbolic versions of the small-numerosity comparison tasks were correlated (Figure 3c) while NDE for RT on the two versions of the tasks were not (Figure 3d).

It has been recently suggested that the NDE may be a problematic measure since it does not directly assess the precision of the ANS and the size of an individual's NDE will depend on the particular ratios or numerical distances involved (Gilmore, et al., 2010; in review). Therefore we also examined the relationships between the small-numerosity comparison task and other tasks using mean accuracy as the measure of performance. These analyses replicated those involving the NDE measure. Participants' mean accuracy on the non-symbolic small-numerosity comparison task was correlated with accuracy on the non-symbolic large-numerosity comparison task ($pr = .555, p < .001$), but not with accuracy on the non-symbolic addition task ($pr = .115, ns$). Finally, using the accuracy measure, performance on the non-symbolic and symbolic versions of the small-numerosity comparison task were correlated ($pr = .314, p = .018$).

Discussion

We found that participants' performance on a range of tasks typically used to assess the ANS are uncorrelated. Estimates of the acuity of the ANS generated from different tasks were therefore unrelated. This is despite the fact that participants' performances on each individual task were in-line with theoretical predictions and

previous findings. These results highlight that there may be significant methodological flaws in tasks typically used to measure the ANS and call into question the claim that individuals use a single system to complete each of these tasks.

The symbolic and non-symbolic comparison and addition tasks used were based on tasks previously used by researchers in this field (e.g. Barth et al., 2003; 2005; Gilmore et al., 2007; Holloway & Ansari, 2009; Pica et al., 2004). Participants' performance on each task showed the expected ratio and distance effects that are characteristic of the ANS and therefore replicated previous findings. Thus, the low correlations found cannot be explained by atypical performance on a particular task or tasks. Most strikingly, estimates of the Weber fraction generated by the non-symbolic large-numerosity comparison and addition tasks were unrelated ($pr = -.09$) even though these tasks involved identical comparison pairs and utilized many of the same stimuli. Furthermore, we examined performance using multiple measures (accuracy, RT, and by generating estimates of the acuity of the ANS where possible) but the lack of correlations was largely consistent and thus cannot be explained by use of any particular outcome measures (cf. Gilmore et al., 2010).

What might account for this pattern of results? There are three possible explanations: the psychometric properties of these tasks may be poor; performance on each task may rely on different sets of domain-general skills; or these tasks may not be solved using a single underlying system.

Little evidence exists concerning the psychometric properties of tasks employed to measure the ANS. Split-half reliability analyses of the tasks used here found satisfactory reliability. However, previous estimates of reliability across two blocks of the same task have provided mixed evidence. Maloney et al. (2010) reported that estimates of participants' NDE was reliable for non-symbolic comparison tasks but not for symbolic comparison tasks. Sasanguie, Defever, Van den Bussche & Reynvoet (in review) found that reliability of NDE estimates on a non-symbolic comparison task were

acceptable when the NDE was calculated on the basis of RT, but not for accuracy. These studies both used small-numerosity comparison tasks and the NDE as the measure of performance. Using Arabic numerals as stimuli, Schneider, Grabner and Paetsch (2009) found that two estimates of individuals' NDEs taken six days apart were relatively weakly correlated ($r=.501$). To our knowledge, reliability of measures of ANS acuity on other ANS tasks and using other measures of performance have not been previously reported. Moreover, participants' performance on these tasks when tested on different occasions separated by days, weeks or months has yet to be tested. If these measures are to be used in future research then it is vital that the psychometric properties of these measures are examined and found to be acceptable.

An alternative explanation of our findings may be that while each task does rely on the ANS, individual tasks also draw on different sets of domain-general abilities and the influence of these other abilities are sufficient to mask the influence of the ANS when examining performance across tasks. Although the tasks used here did utilize many of the same stimuli and methods, there were differences across tasks. For example, the stimuli in the non-symbolic addition task were presented sequentially while in other tasks the stimuli were presented simultaneously and thus the working memory demands of the tasks differ. Also, the non-symbolic versions of the task would pose different types of visuo-spatial demands than the symbolic versions. As intelligence was controlled for in all correlations, however, any domain-general ability that could account for the lack of correlations must also not be strongly correlated with performance on the non-verbal intelligence task. Furthermore, as performance on each task revealed the expected ratio or numerical distance effects predicted by models of the ANS, it is unlikely that these domain-general factors account for more variance in task scores than ANS acuity and explain the lack of correlations across tasks.

Finally, if the psychometric properties of these tests are found to be adequate, and participants' performance is not driven purely by factors such as working memory,

then it is possible that participants are not using a single number system to solve each of these tasks. The hypothesis that participants use the same underlying system to solve this range of tasks stems largely from the common ratio or numerical distance effects that are observed on each task. It is possible, however, that rather than arising from the approximate nature of ANS representations, these effects arise from other sources, such as the comparison process common to all tasks (Van Opstal, Gevers, De Moor, & Verguts, 2008). Maloney et al., (2010) suggested that NDEs elicited by symbolic and non-symbolic versions of the small-numerosity comparison tasks arise from different mechanisms. Recent evidence has also begun to demonstrate that distance effects observed on comparison tasks are unrelated to distance effects elicited from a wider range of tasks that were also previously thought to be due to the approximate nature of ANS representations. Sasanguie et al., (in review) found that distance effects elicited by a comparison task and a priming task were unrelated, even though both had been assumed to arise from approximate overlapping non-symbolic representations.

If performance across these tasks does not arise from a single mechanism, what can explain both the similarities and differences in performance characteristics? Verguts and Fias (2004) suggested that the input to magnitude pathways for non-symbolic and symbolic stimuli differ, and thus although both map to the same underlying magnitude representations, the precision of the magnitude representation will differ according to the input characteristics. In contrast, Cohen Kadosh & Walsh (2009) proposed that there is no single abstract magnitude representation and thus symbolic and non-symbolic tasks will draw on separate systems of representations. The evidence presented here suggests that not only may ratio and distance effects on symbolic and non-symbolic tasks arise from different mechanisms, but that variations in non-symbolic tasks (e.g. comparison compared to addition) may also draw on different systems. Future research is needed to reveal the extent to which performance on a range of ANS tasks reflect the use of a single, or multiple systems of numerical representation and processing.

Given our findings it is unsurprising that previous research has found mixed evidence concerning the relationship of the ANS to formal mathematical skills (e.g. Halberda et al., 2008; Iuculano et al., 2008). On the basis of evidence in favor of a relationship between ANS acuity and learning mathematics, it has been suggested that educational approaches may be designed that draw on this underlying system (e.g. Barth et al., 2005; Gilmore et al., 2007). However, until we are able to reliably measure the acuity of an individual's ANS and demonstrate that a single system does underpin performance on a range of tasks, these calls may be premature.

Notes

1. Including all participants in the analysis produced an essentially identical pattern of results, albeit with somewhat lower correlation coefficients due to the increased noise.
2. Excluding IQ as a covariate produced essentially the same pattern of results for all analyses after Bonferroni corrections.

References

- Barth, H., La Mont, K., Lipton, J., Dehaene, S., Kanwisher, N., & Spelke, E., (2006). Non-symbolic arithmetic in adults and young children. *Cognition*, 98, 199-222.
- Barth, H., La Mont, K., Lipton, J., & Spelke, E.S. (2005). Abstract number and arithmetic in preschool children. *Proceedings of the National Academy of Sciences*, 102, 14116-14121.
- Barth, H., Kanwisher, N., & Spelke, E.S. (2003). The construction of large number representations in adults. *Cognition*, 86, 201-221.
- Cantlon, J.F., Cordes, S., Libertus, M.E., & Brannon, M., (2009). Comment on “Log or linear? Distinct intuitions of the number scale in Western and Amazonian indigene cultures”. *Science*, 323, 38b.
- Cohen Kadosh, R., & Walsh, V. (2009). Numerical representation in the parietal lobes: Abstract or not abstract? *Behavioral and Brain Sciences*, 32, 313-373.
- Cordes, S., Gelman, R., Gallistel, C.R., & Whalen, J. (2001). Variability signatures distinguish verbal from nonverbal counting for both large and small numbers. *Psychonomic Bulletin & Review*, 8, 698-707.
- Dehaene, S. (1997). *The number sense*. Oxford, UK: Oxford University Press.
- Dehaene, S. & Akhavein, R. (1995). Attention, automaticity, and levels of representation in number processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 314-326.
- Dehaen, S., Dupoux, E., & Mehler, J. (1990). Is numerical comparison digital? Analogical and symbolic effects in two-digit number comparison. *Journal of Experimental Psychology: Human Perception and Performance*, 16, 626-641.
- Dehaen, S., Izard, V., Spelke, E., & Pica, P. (2008). Log or linear? Distinct intuitions of the number scale in Western and Amazonian indigene cultures. *Science*, 320, 1217-1220.
- Feigenson, L., Dehaene, S., & Spelke, E.S. (2004). Core systems of number. *Trends in*

Cognitive Sciences, 8, 307-314.

- Flombaum, J.I., Junge, J.A. & Hauser, M.D. (2005). Rhesus monkeys (*Macaca mulatta*) spontaneously compute addition operations over large numbers. *Cognition*, 97, 315-325.
- Gilmore, C.K., McCarthy, S.E., & Spelke, E.S. (in review). Relationships between symbolic and non-symbolic arithmetic in 5-year-old children. *Manuscript submitted for publication*.
- Gilmore, C.K., McCarthy, S.E., & Spelke, E.S. (2007). Symbolic arithmetic knowledge without instruction. *Nature*, 447, 589-591.
- 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, 394-406.
- Halberda, J., Mazocco, M.M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, 455, 665-668.
- Holloway, I.D., & Ansari, D. (2008). Domain-specific and domain-general changes in children's development of number comparison. *Developmental Science*, 11, 644-649.
- 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, 17-29.
- Iuculano, T., Tang, J., Hall, C.W.B., & Butterworth, B. (2008). Core information processing deficits in developmental dyscalculia and low numeracy. *Developmental Science*, 11, 669-680.
- Maloney, E.A., Risko, E.F., Preston, F., Ansari, D., & Fugelsang, J. (2010). Challenging the reliability and validity of cognitive measures: The case of the numerical distance effect. *Acta Psychologica*, 134, 154-161.
- Meck, W.H., & Church, R.M. (1983). A Mode Control Model of Counting and Timing

- Processes. *Journal of Experimental Psychology: Animal Behavior Processes*, 9, 320-334.
- Moyer, R.S., & Landauer, T.K. (1967). Time required for judgements of numerical inequality. *Nature*, 215, 1519-1520.
- Pica, P., Lemer, C., Izard, V., & Dehaene, S. (2004). Exact and approximate arithmetic in an Amazonian indigene group. *Science*, 306, 499-503.
- Sasanguie, D., Defever, E., Van den Bussche, E., & Reynvoet, B. (in review). The reliability of and the relation between numerical distance effects in comparison, same-different judgments and priming. *Manuscript submitted for publication*.
- Van Opstal, F., Gevers, W., De Moor, W. & Verguts, T. (2008). Dissecting the symbolic distance effect: comparison and priming distance effects in numerical and nonnumerical orders. *Psychonomic Bulletin & Review*, 15, 419-425.
- Schneider, M., Grabner, R. H., & Paetsch, J. (2009). Mental number line, number estimation, and mathematical achievement: Their interrelations in grades 5 and 6. *Journal of Educational Psychology*, 101, 359-372.
- Temple, E., & Posner, M.I. (1998). Brain mechanisms of quantity are similar in 5-year-old children and adults. *Proceedings of the National Academy of Sciences*, 95, 7836-7841.
- Verguts, T., & Fias, W. (2004). Representation of number in animals and humans: A neural model. *Journal of Cognitive Neuroscience*, 16, 1493-1504.
- Wechsler, D. (1999). Wechsler Abbreviated Scale of Intelligence. Psychological Corporation.
- Xu, F., & Spelke, E.S. (2000). Large number discrimination in 6-month-old infants. *Cognition*, 74, B1-B11.
- Zorzi, M., & Butterworth, B. (1999). A computational model of number comparison. In M. Hahn & S. C. Stoness (Eds.), *Proceedings of the twenty first annual conference of the Cognitive Science Society* pp 772-777, Mahwah, NJ: Lawrence Erlbaum.

Author Note

This research was supported by a grant from the Economic and Social Research Council (C.G. & M.I.), a British Academy Postdoctoral Fellowship (C.G.), and a Royal Society Worshipful Company of Actuaries Research Fellowship (M.I.).

Table 1: Correlations among non-symbolic ANS tasks

Tasks and measures		<i>pr</i>	Significance
Large-numerosity comparison accuracy	Addition accuracy	.044	<i>ns</i>
	Small-numerosity comparison	.500	$p < .001$
	NDE (accuracy)		
	Small-numerosity comparison	.113	<i>ns</i>
	NDE (RT)		
Large-numerosity comparison <i>w</i>	Addition <i>w</i>	-.008	<i>ns</i>
Addition accuracy	Small-numerosity comparison	.145	<i>ns</i>
	NDE (accuracy)		
	Small-numerosity comparison	.197	<i>ns</i>
	NDE (RT)		

Table 2: Correlations between symbolic and non-symbolic versions of each task

Task and measure	<i>pr</i>	Significance
Large-numerosity comparison (accuracy)	-.026	<i>ns</i>
Large-numerosity comparison (w)	.020	<i>ns</i>
Addition (accuracy)	-.060	<i>ns</i>
Addition (w)	.017	<i>ns</i>
Small-numerosity comparison (NDE accuracy)	.307	$p = .021$
Small-numerosity comparison (NDE RT)	.215	<i>ns</i>

Figure Captions

Figure 1: Ratio and numerical distance effects: a) performance on the symbolic and non-symbolic versions of the large-numerosity comparison task (solid lines) and addition task (dotted lines); b) Accuracy (solid lines) and RT (dotted lines) for the symbolic and non-symbolic versions of the small-numerosity comparison task. Error bars show s.e.m.

Figure 2: Correlations among non-symbolic tasks : a) accuracy on large-numerosity comparison accuracy and addition tasks b) accuracy numerical distance effect (NDE) on small-numerosity comparison task and accuracy on large-numerosity comparison task c) accuracy numerical distance effect (NDE) on small-numerosity comparison task and accuracy on addition task. Accuracy scores are standardized residuals after controlling for IQ.

Figure 3: Correlations between symbolic and non-symbolic versions of each task: a) accuracy on large-numerosity comparison tasks b) accuracy on addition tasks c) accuracy numerical distance effect (NDE) on small-numerosity comparison tasks and d) RT NDE on small-numerosity comparison tasks. Accuracy scores are standardized residuals after controlling for IQ.





