Acta Psychologica 140 (2012) 50–57

Contents lists available at SciVerse ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/ locate/actpsy

Nonsymbolic numerical magnitude comparison: Reliability and validity of different task variants and outcome measures, and their relationship to arithmetic achievement in adults

Gavin R. Price *, Daniel Palmer, Christian Battista, Daniel Ansari

Numerical Cognition Laboratory, Department of Psychology, The University of Western Ontario, Canada

ARTICLE INFO

Article history:
Received 28 September 2011
Received in revised form 20 February 2012
Accepted 22 February 2012
Available online 22 March 2012

Keywords:
Numerical ratio effect
Nonsymbolic number comparison
Approximate number sense
Arithmetic fluency
Magnitude comparison

ABSTRACT

The numerical ratio effect (NRE) and the Weber fraction (*w*) are common metrics of the precision of the approximate numbers sense (ANS), a cognitive mechanism suggested to play a role in the development of numerical and arithmetic skills. The task most commonly used to measure the precision of the ANS is the numerical comparison task. Multiple variants of this task have been employed yet it is currently unclear how these affect metrics of ANS acuity, and how these relate to arithmetic achievement. The present study investigates the reliability, validity and relationship to standardized measures of arithmetic fluency of the NRE and *w* elicited by three variants of the nonsymbolic number comparison task. Results reveal that the strengths of the NRE and *w* differ between task variants. Moreover, the reliability and validity of the reaction time NRE and the *w* were generally significant across task variants, although reliability was stronger for *w*. None of the task variants revealed a correlation between ANS metrics and arithmetic fluency in adults. These results reveal important consistencies across nonsymbolic number comparison tasks, indicating a shared cognitive foundation. However, the relationship between ANS acuity and arithmetic performance remains unclear.

1. Introduction

The last twenty years have witnessed a rapid growth of interest in how the human mind represents numbers and numerical information. In particular, much attention has been paid to the concept of an 'approximate number sense' (ANS). The ANS is a cognitive faculty that represents the cardinality of sets of objects in an approximate fashion (Dehaene, 1997; Dehaene, Dehaene-Lambertz, & Cohen, 1998), and is thought to be phylogenetically ancient, shared with human infants and non-human primates (Brannon, 2006; Hubbard et al., 2008). The concept of the ANS has been influential in the field of numerical cognition, and the principle way in which it has been measured is through numerical comparison paradigms (Ansari, 2008; Dehaene & Akhavein, 1995; Halberda, Mazzocco, & Feigenson, 2008). In these tasks, participants are asked to compare the relative numerical magnitude of two numerical stimuli (e.g. which is larger, 7 or 3?). These stimuli can be either nonsymbolic (e.g. a set of dots or squares) or symbolic (e.g. Arabic digits or written number words), and typically, the ANS is thought to be most directly measured through nonsymbolic number comparison, since comparing numerical magnitudes represented in a symbolic format require the additional step of mapping between symbols and the magnitudes they represent. Previous studies have endeavored to link measures of ANS acuity to performance on standardized measures of math achievement (e.g. Halberda et al., 2008; Holloway & Ansari, 2009), yet the results across studies have been inconsistent, and thus the present study investigates possible sources of that inconsistency by contrasting the reliability, validity and relationship to math achievement of two different ANS metrics calculated from three variants of the nonsymbolic numerical magnitude comparison task.

The numerical comparison (both symbolic and nonsymbolic) produces a key metric of the ANS, the 'numerical distance effect' (NDE) (Moyer & Landauer, 1967). The NDE refers to a monotonic increase in both reaction time and error rates as the numerical distance between the two comparators decreases. Thus, individuals are typically faster and more accurate when comparing the numerical magnitude of 5 vs 9 than comparing 8 vs 9.

The NDE is a highly replicable effect, indexed by the difference in reaction time and accuracy of responses to comparisons with relatively small or large distances, or alternatively, by calculating the slope that relates reaction time and accuracy to numerical distance (Dehaene et al., 1998; Maloney, Risko, Preston, Ansari, & Fugelsang, 2010). This effect has been interpreted as a performance product of the noise within the mental representation of numbers (Dehaene & Cohen, 1995), or alternatively, of response resolution processes, common to comparing two stimuli of any type (Van Opstal, Gevers, De Moor, & Verguts, 2008). Researchers have also applied the Weber–

* Corresponding author at: Numerical Cognition Laboratory, Department of Psychology, The University of Western Ontario, London, ON, N6A 2K7, Canada. Tel.: +1 519 661 2111x84596; fax: +1 519 850 2554.
E-mail address: gavinprice@gmail.com (G.R. Price).

0001-6918/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.actpsy.2012.02.008

Fechner Law in order to calculate a metric of the noisiness of an individual's ANS, the so called 'internal Weber fraction'(*w*). Therefore, individuals with less precise ANS representations should have a relatively higher *w* (i.e. more noise within the representation) than individuals with more precise representations, and by extension, a higher *w* should relate to poorer ability to discriminate relative numerical magnitudes.

The performance detriments associated with decreasing numerical distance between comparators can also be described in terms of the ratio between the two numbers to be compared. The 'numerical ratio effect' (NRE) refers to a monotonic increase in reaction time and error rate as the *ratio* (smaller/larger) between the two numbers increases. The NRE and the NDE are highly correlated, and so can be discussed interchangeably when referring to the extant literature. For the sake of clarity, from this point on we refer solely to the NRE.

Particularly in the case of nonsymbolic numerical comparison, multiple variants of the task have been employed (e.g. overlapping sets of dots, spatially separate sets of dots, one set of dots followed by another) (Ansari, Lyons, van Eimeren, & Xu, 2007; Halberda et al., 2008; Holloway & Ansari, 2009), and as discussed above, multiple metrics have been calculated from the resulting data (e.g. numerical distance effect, numerical ratio effect and *W*). Despite this wide variation in paradigm variants, relatively little attention has been paid to the reliability and validity of the resulting dependent variables, and typically, the results of different studies are discussed without reference to the subtle but potentially influential variations in task structure. In other words, there is an assumption that different task variants are all tapping into the same underlying system of representation. This assumption needs to be empirically examined. Furthermore, to date there has been no empirical investigation of within task differences between different ANS metrics (e.g. numerical ratio effect and *W*) in terms of their reliability and validity. Yet, while we know little about the relationship between the different ANS metrics produced by different task paradigms, a growing number of studies are drawing links between these metrics and higher level arithmetic abilities based on the hypothesis that a more precise ANS will facilitate more successful acquisition of arithmetic skills. In order to better understand the significance of relationships (or the lack thereof) between different measures of the ANS and mathematical achievement, it is therefore necessary to better understand the reliability and validity of these different dependent measures of number comparison.

Individual variation in performance on standardized tests of arithmetic has been found to correlate with individual *W*s during a nonsymbolic numerical magnitude comparison task (Halberda et al., 2008; Libertus, Feigenson, & Halberda, 2011; Mazzocco, Feigenson, & Halberda, 2011a, 2011b) and the NRE during symbolic number comparison (Holloway & Ansari, 2009), suggesting that the ANS plays a role in the development of arithmetic skill. Furthermore, children with developmental dyscalculia (DD), a specific arithmetic learning disorder, have been found to show atypical *W*s and NRE's relative to typically developing peers (Mazzocco et al., 2011a, 2011b; Mussolin, Mejias, & Noel, 2010; Price, Holloway, Räsänen, Vesterinen, & Ansari, 2007). However, some inconsistency exists in the current empirical literature on this issue. For example, while Halberda et al. (2008) reported a relationship between ANS acuity (*w*), measured through a nonsymbolic number comparison task, and arithmetic performance, Holloway and Ansari (2009) and Mundy and Gilmore (2009) both reported no such correlation between the NDE during nonsymbolic comparison and standardized math achievement scores.

The source of the apparent contradiction between those studies is unclear, but one possibility is that the different variants of the nonsymbolic comparison used in those studies (e.g. Halberda et al. (2008) employed an overlapping design, in which sets of yellow and blue dots were intermixed in a single display, while Holloway and Ansari (2009), as well as Mundy and Gilmore (2009), employed a paired presentation design, in which two distinct sets of squares were simultaneously presented) differ in their internal reliability, and thus, result in significantly different levels of individual variation in the NRE. Another alternative is that differences in the precision with which the *W* and the NRE respectively reflect the acuity of the ANS result in different results when relating those metrics to standardized arithmetic performance. The reliability and validity of variants of the numerical comparison paradigm and the ANS metrics they yield are two issues that have only recently started to receive attention, but are of great importance if we are to meaningfully compare the results of different studies and evaluate their potential as correlates of individual differences in higher-level numerical and arithmetic skills.

Recently, Maloney et al. (2010) investigated the internal reliability and convergent validity of symbolic and nonsymbolic numerical comparison tasks using two variants of each task. The results indicated that while comparison of two simultaneously presented symbolic and nonsymbolic stimuli showed significant reliability in both reaction time and error, comparing single stimuli to the reference number 5 did not. Convergent validity between symbolic and nonsymbolic tasks was not significant, irrespective of the variant. Convergent validity was significant between the two nonsymbolic but not the symbolic tasks. That lack of validity between conditions is supported by a more recent investigation (Gilmore, Attridge, & Inglis, 2011) reporting that while participants' performance on the nonsymbolic and symbolic tasks showed significant internal reliability, measures of convergent validity between the tasks were mostly non-significant. In a similar study, Sasanguie, Defever, Van den Bussche, and Reynvoet (2010) investigated the reliability of the NRE with and between three non-symbolic numerical processing tasks. Results from that study revealed significant internal reliability for the comparison and same–different tasks, but not for the priming comparison task. Furthermore, the NREs in the comparison and same–different tasks were significantly correlated, but neither correlated with the NRE in the priming comparison task. The authors suggest, therefore, that the comparison and same–different tasks are trustworthy tasks for the measurement of the NRE, whereas the use of the priming comparison task requires caution.

It is apparent from the literature reviewed above that data on the relationship between ANS acuity and arithmetic performance are inconsistent, and only a handful of studies have investigated a potential source of that inconsistency, the reliability and validity of ANS metrics resulting from different numerical comparison task paradigms. If the ANS is to form a cornerstone of cognitive theories of number processing, it is essential to characterize the reliability and validity of numerical comparison and the NRE. This is true for both practical and theoretical reasons. At a practical level, the NRE is rapidly becoming a popular and widely used measure of individual differences in ANS acuity, yet several variants of the task exist, and there is currently no basis to assume that each variant taps into the ANS in the same way and to the same degree.

From a theoretical standpoint, it is important to consider that although the NRE is a highly replicable effect, that replicability does not necessarily imply that in all cases the NRE reflects a direct measure of the acuity of the ANS, free from the influence of additional cognitive demands which may vary between task variants. In order to interpret the NRE in a theoretically meaningful way, we need to understand how its strength, reliability, validity, and relationship to formal measures of arithmetic vary between different variants of the numerical comparison paradigm. The current study addresses these issues by contrasting the relative strength of the NREs derived from three nonsymbolic number comparison task variants. The NRE is a direct measure of an individual's performance on a numerical comparison task, whereas the *W* is an estimation of the noise in the representation that drives that performance and can be used to predict participants' relative magnitude comparison performance for any given ratio. So, while theoretically, *W*s and NREs should be determined by the same cognitive structures, they are, in essence, fundamentally different measures of

the ANS. Therefore, in the present study we calculated individual *W*s as well as NREs, to be able to compare our results directly with previous studies which have used NREs and *W*s separately.

For both the NRE and *w* measures, we computed internal reliability, convergent validity, and importantly, the extent to which these metrics of the ANS correlate with standardized measures of math fluency in adults. These comparisons allow us to test the hypothesis that the apparently contradictory results of previous studies using different variants of the nonsymbolic comparison paradigm discussed are driven by variations in the task design and/or the metrics used to index ANS precision.

2. Methods

2.1. Participants

The present sample comprised 39 undergraduate students enrolled at The University of Western Ontario, Canada. Participants were recruited through upper year undergraduate courses or through the first year psychology resource pool via signup sheets and internet forms. All participants completed all experimental conditions and provided informed consent that was approved by the Psychology Research Ethics Board at the University of Western Ontario. Three participants were excluded from analysis due to accuracy rates below 70%. This threshold was set to ensure that participants were focused on the task, but more importantly, to allow a reasonable estimation of the *W*s for each participant. Consequently, 36 participants constituted the final sample (15 males, 21 females, Mean age 22.29 yrs, Std Dev, 4.49 yrs).

2.2. Materials and procedure

The experiment comprised three variations of nonsymbolic numerical magnitude comparison tasks, paired, sequential, and intermixed. The conditions are described below.

2.2.1. Paired presentation

In the paired presentation condition, illustrated in Fig. 1, two dot arrays were presented simultaneously, one to the left of the screen and one to the right. In half of the trials, blue was the correct answer and in half, yellow was the correct answer. Color and response side were fully counterbalanced. Trials consisted of 750 ms stimulus presentation followed by 500 ms presentation of a visual mask, followed by 750 ms presentation of a fixation cross. Participants were required to select via button press whether the set on the right or the left contained more dots. The number of dots in each of the arrays ranged from 6 to 40.

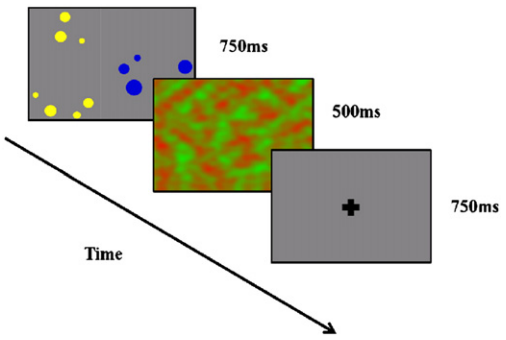

Fig. 1. Paired presentation condition trial schematic.

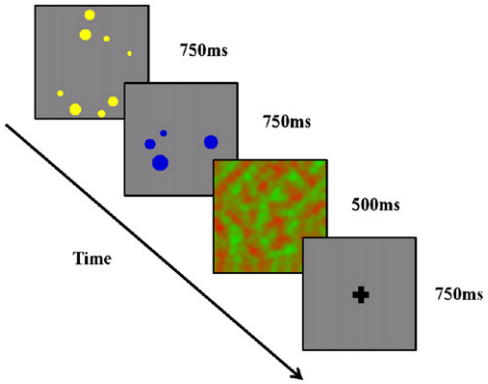

Fig. 2. Sequential presentation condition trial schematic.

2.2.2. Sequential presentation

In the sequential presentation condition, illustrated in Fig. 2, dot arrays were presented in isolation, one after the other. Trials consisted of the first dot array presented for 750 ms followed by the second dot array presented for 750 ms followed by 500 ms presentation of a visual mask, followed by 750 ms presentation of a fixation cross. Participants were required to select via button press whether the blue or yellow array contained more dots. The number of dots in each of the arrays ranged from 6 to 40.

2.2.3. Intermixed presentation

In the intermixed presentation condition, illustrated in Fig. 3, blue and yellow dot arrays were presented as a single large array in the center of the screen. The blue dots and yellow dots were arranged in a non-overlapping (i.e. each dot was spatially distinct from each other dot) random distribution. Trials consisted of 750 ms stimulus presentation followed by 500 ms presentation of a visual mask, followed by 750 ms presentation of a fixation cross. Participants were required to indicate via button press whether there were more blue or more yellow dots within the array. The number of blue and yellow dots in the array ranged from 6 to 40.

For all conditions, color and response side were fully counterbalanced. The fixation cross was added at the end of the trial simply as a function of the e-prime coding structure. Subjectively, participants saw a long initial fixation, followed by a trial, followed by fixation, followed by a trial, and so on. Thus from their perspective each trial was bounded on either side by fixation.

Stimuli were created using custom scripts using the Python programming language. Stimulus presentation and collection of participant

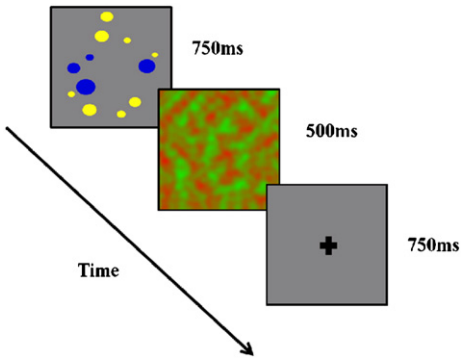

Fig. 3. Intermixed presentation condition trial schematic.

response was also done using custom scripts using the Python programming language (Straw, 2008).

To ensure that total area or the total perimeter of dots was not predictive of numerosity, we matched half of the dot arrays for total area and half of the dot arrays for total perimeter. In other words, 50% of trials had equivalent area while 50% had equivalent perimeter. When perimeter was matched, the array with more dots occupied more cumulative surface area. When cumulative surface area was matched, the array with more dots occupied a greater perimeter. These area-matched and perimeter-matched trials were randomly presented in each presentation condition, in an effort to prevent the participant from developing a strategy that relied on the relative size of the dot arrays. The average size of the dot arrays was 12.5 cm

Trials were classified according to ratio based on the ratio of the larger set to the smaller set. The six ratios used were 0.25, 0.33, 0.50, 0.66, 0.75, and 0.90. In this method, a smaller ratio generally equates to a larger absolute numerical distance, so that 0.25 was the easiest ratio condition, progressing through to 0.90 which was the hardest. Each presentation condition included exactly the same distribution of trials, so that the number of trials per ratio, per condition was identical. There were 40 trials per ratio for each condition resulting in a total of 240 trials per presentation condition and 720 trials in total. Additionally, there was a break between every 60 trials to produce 4 blocks for each presentation type.

The visual mask was a large square that contained a blend of green and red (see Figs. 1–3). Masking was carried out to effectively eradicate after images of the blue and yellow dot patterns so that responses could not be made on the basis of after images in the case of the sequential condition, nor interfere with subsequent trials in the paired and intermixed conditions. Responses were accepted during the stimulus, mask and fixation presentation, but any responses made after the fixation were coded as incorrect. In addition, during the sequential condition, any responses made prior to the appearance of the second dot array were coded as incorrect. Participants responded to trials by pressing the left and right control keys on the laptop used for presentation. In the paired condition the left key was pressed if the left array was larger, and the right key if the right array was larger. In the sequential and intermixed conditions the left key was pressed if the yellow array contained more dots, and the right key was pressed if the blue array contained more dots. Color was counterbalanced across response keys.

3. Results: numerical ratio effect

For all analyses, the NRE was calculated by computing the unstandardized coefficient beta of a linear regression analysis for each individual subject. This slope value was then used as the dependent variable to quantify the NRE for both reaction time and accuracy analyses. One sample t-tests conducted for reaction time and accuracy slopes for each presentation condition revealed a significant slope in all cases. In other words, the NRE was significant for both accuracy and reaction time for each presentation condition.

Table 1
mean reaction time and accuracy values, slopes and Weber fractions for each presentation condition.

	Intermixed		Paired		Sequential	
	Mean	SD	Mean	SD	Mean	SD
Reaction time	961.16	284.92	1024.71	226.87	976.49	296.52
Accuracy	0.82	0.05	0.89	0.22	0.89	0.04
Reaction time slope	188.05	121.28	314.72	164.92	256.2	144.95
Accuracy slope	−0.50	0.10	−.055	0.09	−0.43	0.11
Weber fraction	0.38	0.13	0.22	0.04	0.21	0.06

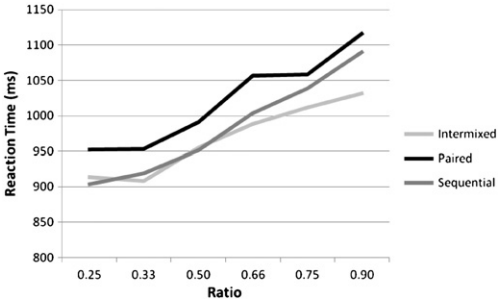

Fig. 4. Mean reaction time slopes for each presentation condition.

Table 1 shows mean values for reaction time and accuracy, mean slopes for reaction time and accuracy, and mean 'w' for each presentation condition.

3.1. Presentation condition and the numerical ratio effect

3.1.1. Reaction time

To investigate the relative strength of the numerical ratio effect (NRE) in reaction time between presentation conditions, we carried out a one-way analysis of variance (ANOVA), contrasting the strength of the reaction time slope between conditions (mean reaction times slopes are presented in Fig. 4).

This analysis revealed a significant main effect of presentation format ($F(2, 105) = 6.9$, $p < .01$) on the strength of the reaction time slope. Posthoc Tukey's HSD tests revealed the strongest RT slope in the paired presentation condition, which was significantly stronger than the intermixed condition RT slope (Mean Difference = 126.67, Std Error = 34.13, p = .001), but not significantly different from the sequential condition RT slope (Mean Difference = 58.52, Std Error = 34.13, p = .21). The intermixed and sequential conditions did not differ significantly from one another in the strength of their RT slopes (Mean Difference = 68.14, Std Error = 34.13, p = .12).

3.1.2. Accuracy

The slopes of response accuracy, presented in Fig. 5, were clearly not linear, and thus using linear regression to analyze these data was not appropriate. Therefore we did not conduct equivalent analyses on the accuracy data to those conducted on the reaction time data. Accuracy analyses are instead carried out using the calculation of *w*, as reported below.

3.2. Reliability and validity analyses

To examine the reliability of the NRE, the slope from the first half (first 120 trials) of the trials for each condition was correlated with the slope from the second half of the trials (second 120).

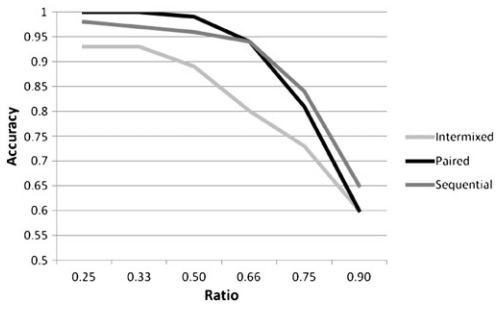

Fig. 5. Mean accuracy slopes for each presentation condition.

Table 2
Correlation values for reaction time NREs (slopes) and math fluency standard scores.

	Reaction time (millseconds)
Intermixed	r(34) = −.03, n.s.
Paired	r(34) = .01, n.s.
Sequential	r(34) = .01, n.s.

3.2.1. Reliability

For the reaction time NRE, significant reliability was revealed for the intermixed ($r\ (34) = .57$, $p < .001$), paired ($r\ (34) = .78$, $p < .001$) and sequential ($r\ (38) = .65$, $p < .001$) presentation conditions.

3.2.2. Validity

To calculate convergent validity, reaction time and accuracy slopes were calculated from all trials for each presentation condition (i.e. without separating it by halves). These slopes were then correlated between presentation formats to calculate the validity between presentation formats.

The results of these analyses reveal that for the reaction time NRE, only the intermixed and paired conditions showed significant validity ($r(34) = .52$, $p = .001$). Intermixed and sequential conditions showed a marginally significant correlation ($r\ (34) = .32$, $p = .055$). The paired and sequential conditions revealed a nonsignificant trend towards a correlation ($r\ (34) = .29$, $p = .09$).

3.3. The NRE and arithmetic fluency

To investigate any differences between presentation conditions in the relationship between the NRE and standardized measures of arithmetic, we correlated individual reaction time slopes from each presentation condition with standard scores on the Woodcock Johnson Math Fluency subtest (Woodcock, McGrew, & Mather, 2001). The mean standard score for the current sample was 105.42 (standard deviation 16.1, range 64–140).

The results of these analyses, summarized in Table 2, revealed no association between the NRE for reaction time and math fluency scores, in any of the presentation conditions.

4. Results: Weber fraction

To calculate an internal 'w' for each participant, we sought to obtain a value for w such that for each numerosity pairing (given by n1 and n2), the expected error rate produced by the formula below was as close as possible to the error rate observed in the data. In other words, we fit the model below (Fig. 6) to the data, using *w* as the free term.

This fitting employs the following strategy (for each participant and/or condition of interest).

1. Obtain the n1/n2 pairings and the error rates for each.
2. For each possible value of w (we use values of w between 0 and 1 in steps of 0.01), calculate the predicted error rate using the above equation.
3. For each predicted error rate (for all the n1/n2 pairings), compare that to the error rate observed in the data. For one value of w, we will have an observed and expected error rate for each n1/n2 pairing.
4. Calculate the fitness of the model (for each w) by calculating the sum of squared differences between the observed and expected differences.
5. Select the w with the best fitness (i.e., lowest sum of squared difference).

$$\frac{1}{2}\operatorname{erfc}\left(\frac{n_1 - n_2}{\sqrt{2}w\sqrt{n_1^2 + n_2^2}}\right)$$

Fig. 6. Model for calculating Weber fraction.

4.1. Effect of presentation condition

To investigate the relative strength of the 'w' between presentation conditions, we carried out a one-way analysis of variance (ANOVA), comparing mean *w* between conditions.

This analysis revealed a significant main effect of presentation condition ($F(2, 105) = 45.63$, $p < .001$). Posthoc Tukey's HSD Tests revealed that the intermixed condition yielded a significantly higher mean *w* than the paired (Mean Difference = .164, Std Error = .02, $p < .001$) and sequential (Mean Difference = .173, Std Error = .02, $p < .001$) conditions. The paired and sequential conditions, on the other hand, did not differ from one another (Mean Difference = .01, Std Error = .02, *n.s.*).

4.2. Reliability and validity analyses

To examine the reliability of the 'w', the *w* from the first half (first 120) of the trials for each condition was correlated with the *w* from the second half of the trials (second 120).

4.2.1. Reliability

The results of this analysis revealed significant reliability for the intermixed ($r\ (34) = .78$, $p < .001$), paired ($r\ (34) = .47$, $p < .005$) and sequential ($r\ (34) = .44$, $p < .01$) presentation conditions.

4.2.2. Validity

To calculate convergent validity of the 'w' between conditions, *w* scores for each condition were correlated with each other.

The results of these analyses, reveal significant validity between the intermixed and paired conditions ($r\ (34) = .39$, $p < .05$), the intermixed and sequential conditions ($r\ (34) = .68$, $p < .001$), and between the paired and sequential conditions ($r\ (34) = .50$, $p < .01$). In other words, the *w* scores for each condition correlated significantly with the *w* scores for each other condition.

4.3. The Weber fraction and arithmetic fluency

To investigate any differences between presentation conditions in the relationship between the 'w' and standardized measures of arithmetic, we correlated individual *w* values from each presentation condition with standard scores on the Woodcock Johnson Math Fluency subtest (Woodcock et al., 2001).

The results of these analyses, revealed no association between the *W* and math fluency scores, in the intermixed (r(34) = −.24, n.s.), sequential (r(34) = −.1, n.s.) or paired (r(34) = −.28) presentation conditions.

4.4. Correlation between reaction time NRE and w

To further characterize the relationship between the NRE calculated from reaction time slopes and individual w's, calculated from response accuracy, we correlated NRE and w for each condition. This analysis revealed a significant correlation between NRE and w in the sequential condition (r(34) = −.331, p = .048). No other correlations were significant. This suggests that the relationship between RT slope and w is not very strong, which might be explained by the fact that one is a measure of RT while the other is a measure of accuracy.

5. Discussion

The numerical ratio effect is a common and highly replicable psychophysical product of comparing relative numerosities. These effects are widely interpreted as metrics of a mental representation of numbers organized along a 'mental number line'. An alternative index to the NRE is the Weber fraction (w), an estimate of the noise associated this representation, rather than a behavioral response profile. The precision of this representation is thought to reflect the acuity of the approximate number sense (ANS), a foundational cognitive capacity suggested to play an important role in the development of numerical and arithmetic skills. However, evidence linking the acuity of the ANS to performance on standardized measures of arithmetic ability is inconsistent across studies (Halberda et al., 2008; Holloway & Ansari, 2009; Inglis, Attridge, Batchelor, & Gilmore, 2011). The present study sought to shed light on the sources of this inconsistency by comparing the size of the NRE and *W*s elicited by three variants of the nonsymbolic number comparison task, and contrasting their internal reliability, convergent validity and relationship to standardized measures of arithmetic fluency.

The results of this study revealed that while all non-symbolic number comparison conditions showed a significant NRE, the paired presentation condition elicited the strongest NRE for both reaction time and accuracy. The intermixed condition yielded the weakest reaction time NRE while the sequential condition yielded the weakest accuracy NRE. These results suggest that different variants of the nonsymbolic number comparison paradigm are liable to produce significantly different strengths of the NRE. We speculate that the difference in the strengths of the NRE between conditions is driven by the influence of extraneous domain general cognitive demands. In other words, we do not suggest that the three variants differ in the extent to which they index the ANS, but rather that the clarity of that index is influenced by additional cognitive processing demands such as working memory in the sequential condition, and visual resolution of overlapping comparators in the intermixed condition. Logically, the paired presentation condition requires the least extraneous cognitive processing because the two displays are spatially separate, and so no visual segmentation is required. Furthermore, the simultaneous presentation negates any working memory demands. In keeping with this, we see the strongest NRE elicited in that variant. However, it should also be noted that whether a paradigm that produces a stronger NRE is in some way a better measure of the underlying ANS is difficult to say. It could be argued at least, that by eliciting the strongest mean NRE across the sample, the paired presentation variant may be the most robust method by which to extract significant NREs from all subjects.

When contrasting the size of the average *W* between conditions, as opposed to the size of the NRE, significant differences between conditions were also found. In this case, however, the intermixed presentation condition elicited the highest mean *w*, followed by the paired and sequential conditions, whose mean *w* did not differ from one another. The fact that the *W* varies between conditions is of interest because, as with the NRE, it suggests that this metric cannot be taken as an isomorphic representation of the ANS, independent of task features. Also of interest is that the condition that produced the highest NRE slope (paired) was not the same condition that produced the highest mean *w* (intermixed). This is most likely due to the fact that the *w* is influenced by overall error rate, which was highest in the intermixed condition, whereas the NRE slopes are dependent more on relative changes in accuracy or reaction time. It is important to note, given the variation in the NRE and *w* between conditions, the absolute values of these ANS metrics are likely to offer limited information regarding the nature of the underlying representation they are thought to reflect. While group differences in such measures and relationships to standardized achievement measures may be informative, as isolated values their absolute magnitude is of limited utility.

When testing the reliability of the NRE for each condition, it was observed that all conditions showed significant internal reliability for the reaction time NRE. As discussed above, each of the three variants tested in the current study introduce different degrees and types of domain general cognitive demands, which may affect the strength of the NRE or *w* elicited by each variant. However, the results of the internal reliability analyses suggest that these extraneous factors do not appear to negatively affect the reliability of the reaction time NRE. When using *w*'s as opposed to NRE slopes, the results of the reliability analysis mirrored those for the reaction time NRE, in that all conditions demonstrated significant reliability. These results suggest that, disregarding the accuracy NRE, neither the presentation format nor the calculated metric of the ANS undermine the reliability of the obtained effects, and support the results of Maloney et al. (2010), Sasanguie et al. (2010) and Gilmore et al. (2011) who all reported significant reliability for a range of nonsymbolic comparison tasks.

With regards to convergent validity, only the paired and intermixed conditions revealed a significant correlation between reaction time NREs. The intermixed and sequential conditions revealed marginally significant validity ($p = .055$), while the paired and sequential conditions showed a nonsignificant trend toward validity ($p = .09$). Similarly, when correlating *W*s, all conditions showed significant validity with each other. The consistent convergent validity between number comparison paradigms in the present study supports the results of Maloney et al. (2010) and Sasanguie et al. (2010), who observed significant validity between two nonsymbolic comparison paradigm variants and nosymbolic comparison and same/different judgements respectively. Gilmore et al. (2011), on the other hand, did not observe such consistent validity between nonsymbolic ANS measures, and therefore suggest that the absence of convergent validity may be an indication that different cognitive systems are being employed to solve different comparison tasks, as opposed to drawing on a single abstract ANS. However, in the present study, consistent validity was observed for both the reaction time NRE and *W*s, and, therefore, we suggest that despite the variations in additional cognitive processing imposed by the different presentation conditions, at some level, each variant must be measuring the same core cognitive process.

On the basis of these results then, we suggest that the linear slope of response reaction time is a generally reliable and valid measure across paradigm variants, and thus can be assumed to reflect the acuity of the ANS at some level, regardless of task variations. The *w* also provides a reliable and valid measure of ANS acuity. Reaction time NREs and w's did not correlate with another, except in the case of the sequential task variant, suggesting that the relationship between RT slope and w is not very strong, possibly because one is a measure of RT while the other is a measure of accuracy. This also suggests that if a paradigm is intended to elucidate differences in response accuracy, then w is the preferable dependent variable, while in paradigms with longer exposure time, which might be expected to reveal variability in RT, then RT slope is the preferable dependent variable.

These results suggest that when investigating the effect of ratio on reaction time, the linear slope is a reasonable measure to use, while the *W* provides a reasonable measure when investigating the effects of ratio on response accuracy. However, as noted above, the differences between paradigm variants in the magnitudes of these measures suggests that first, additional cognitive processes are being imposed by the task structure, and second, that the absolute values of the NRE or *w* in of themselves offer limited information as to the nature or characteristics of the mental representation of numerical magnitudes. In fact, the consistent reliability and validity across presentation variations also do not provide information regarding the characteristics of any underlying ANS representation, nor do they necessarily suggest that these tasks are all measuring the ANS with reasonable equivalence. It could be the case that what the data reflect are, instead, consistencies associated with domain general comparison and response selection processes that

cannot be disentangled from the underlying representation in the context of tasks requiring a comparative behavioral response.

Given that reliability and validity was consistent across paradigm variants and ANS metrics, the inconsistent results linking ANS acuity to arithmetic performance remain puzzling. With this consideration in mind, we investigated the extent to which the NREs and *W*s elicited by each of the three variants employed in the present study correlated with individual scores on a standardized test of arithmetic fluency. None of the three variants' NREs or w's correlated with math fluency, suggesting that the relationship between these measures of the ANS and arithmetic skill is far from clear. These results support those reported by Holloway and Ansari (2009), who showed no relationship between the nonsymbolic NDE and math fluency, and those of Inglis et al. (2011), who showed no association between the *w* and arithmetic performance in adults (although that study did reveal such an association in children). On the other hand, the present results contradict those of Halberda et al. (2008) who reported a significant correlation between ANS acuity and arithmetic achievement using a nonsymbolic number comparison task of the intermixed variant used in the present study, as well as those of Libertus et al. (2011) and Mazzocco et al. (2011a, 2011b) who each reported an association between measures of ANS acuity collected from preschool children and their later arithmetic performance.

Several possible explanations may account for this apparent discrepancy. First, Halberda et al. (2008) and Libertus et al. (2011) tested larger samples of participants than the present study, which may have resulted in a greater range of individual differences, subsequently revealing more fine grained relationships between ANS acuity and arithmetic skill. However, Holloway and Ansari (2009) tested 87 participants (more than twice the number of subjects in the present study) and, as in the present study, found no relationship between the NRE using a paired variant of the nonsymbolic comparison task and standardized arithmetic achievement scores, and so, sample size alone cannot explain the absence of correlation in the present study. Second, the discrepancy between the results of Halberda et al. and those of Holloway and Ansari and the present study might be attributed to the specific standardized arithmetic test used. While the present study correlated the NRE with standard scores on the Woodcock Johnson Math Fluency subtest, Halberda et al. correlated their ANS acuity measure with standardized scores on the Woodcock Johnson Calculation (WJC) subtest as well as the 'test of early mathematical ability, second edition' (TEMA-2), while Libertus et al. (2011) and Mazzocco et al. (2011a, 2011b) used the TEMA-3 test. It is possible that the more composite measures, which include a wider variety of items than the WCJ Fluency subtest alone, may have captured a wider variation of performance than was captured in the present study. However, Holloway and Ansari (2009) correlated their measure of ANS acuity (i.e. the numerical distance effect) with both the fluency and calculation subtests of the Woodcock Johnson Math Test, as well as a composite measure of the two and, in all cases, the nonsymbolic NRE did not correlate with standard scores. Furthermore, Inglis et al. (2011) correlated *W*s with the WJC Calculation subtest, and found a correlation in children but not adults. These findings suggest that the absence of a correlation in the present study cannot be attributed to the type of standardized math test alone.

A third possible explanation of the discrepancies between the results of the present study and those of Halberda et al. (2008) is the nature of the ANS acuity metric. Holloway and Ansari (2009) used the numerical distance effect (NDE), while Halberda et al. (2008) calculated '*W*s' for each participant. While the NRE reflects the proportional performance change as a function of numerical ratio or distance, the *w* is a value that estimates the degree of imprecision around the response to a given numerosity comparison (and thus, a lower *w* reflects greater ANS acuity). However, in the present study we correlated both the NRE and *W*s with math fluency standard scores and found no relationship for either, suggesting that the ANS metric used cannot account for the inconsistent findings across studies.

The potential sources of discrepancy discussed above all relate to experimental variables, and do not seem to adequately account for the inconsistent results across studies. It is also possible, however, that the variation in results stems from developmental differences in the ANS. Halberda et al. (2008) used the *W* calculated at age 14 and correlated that retrospectively with standardized math scores collected from ages 5 to 11. Libertus et al. (2011) and Mazzocco et al. (2011a, 2011b) correlated ANS measures collected in preschool with later arithmetic performance. The current study calculated the NRE in participants with an average age of 21.58 years, and correlated that with math fluency scores collected in the same testing session. Inglis et al. (2011) have recently demonstrated that the *W* correlates with standardized calculation scores in a sample of children aged 7–9, but not in a sample of adults aged 18–48. This pattern of results may explain the discrepancies between the present results and those of Halberda et al. (2008). It is possible that the role of the ANS in the development of arithmetic competency is foundational rather than continuous, and thus, the relationship between the two measures might be strong early in the development process, but begins to decline as formal arithmetic abilities become increasingly independent, relying more on enculturated symbolic processing mechanisms (Ansari, 2008). However, this explanation does not account for the differences between the results of Halberda et al. (2008) and Holloway and Ansari (2009). Holloway and Ansari calculated the NDE in children aged 6, 7, and 8, and correlated that with standardized arithmetic scores collected in the same testing session. Thus, while the relationship between ANS acuity and arithmetic ability may change over the course of the development, that change cannot be the only explanation for the inconsistencies in the findings discussed above.

It is also possible that the relationship between the ANS and formal arithmetic varies with arithmetic achievement level as opposed to chronological age. While the present study, and the previous studies reviewed above reveal inconsistent results with regard to participants in the normal range of arithmetic ability, several studies have revealed deficits in nonsymbolic numerical magnitude processing in children with developmental dyscalculia (DD), a specific arithmetic learning disorder (Landerl, Fussenegger, Moll, & Willburger, 2009; Mussolin et al., 2010; Piazza et al., 2010; Price et al., 2007). Furthermore, recent evidence suggests that it is only the more severe cases of arithmetic learning disorders, and not mildly below average difficulties, in which measures of the ANS are found to be atypical (Mazzocco et al., 2011a). Thus, the relationship between ANS acuity and formal arithmetic may be strongest in the lowest performing individuals, however, this relationship and its causes require further empirical investigation in order to be elucidated. Furthermore, it should be noted that the current sample was selected from a fairly narrow demographic (undergraduate students). It is possible that a relationship between the ANS metrics and math fluency would be revealed in a larger and more diverse sample, and is something that future research should address.

In summary, the results of the present study reveal differences in the strength of the NRE and *W* elicited by three variants of the nonsymbolic number comparison paradigm. However, the variant eliciting the strongest effect varied depending on whether we consider the NRE of the *W*, suggesting that the absolute magnitude of these effects is variable between conditions and the nature of the metric, and is, therefore, of limited informative value. We observed consistent reliability and validity across and between paradigm variants, suggesting that the different paradigms tested reliably measure similar, if not the same, cognitive construct, but whether that construct is the acuity of the ANS or simply general comparison and response processes is open to debate. The accuracy NRE was not reliable or valid

in any variant, and thus, does not appear to be an appropriate metric of ANS acuity. Although the reaction time NRE was generally reliable and valid, the correlations within and between conditions observed for the *W*s were stronger, and thus, it appears as though the *W* is the optimal metric for measuring ANS acuity. Finally, we observed no relationship between the NRE or the *W* and standard scores of math fluency for any of the three variants. This suggests that further investigation is required to elucidate the relationship between the ANS and arithmetic skill, and, in particular, under what circumstances, at what age and with what populations such relationships are obtained. However, as the present data reveal, such future investigations may proceed with faith in the measures which they use to index the ANS. Both the numerical ratio effect and the *W* are reliable and valid outcome measures of non-symbolic numerical magnitude comparison regardless of variations in the comparison task structure, and thus, it appears that they do measure a shared core cognitive mechanism, independent of those cognitive processes that vary between task variants.

Acknowledgments

This research was supported by funding from Canadian Institutes of Health Research (CIHR) Operating Grant and Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant to DA, and Ontario Ministry of Research and Innovation (OMRI) Postdoctoral Fellowship to GRP.

References

Ansari, D. (2008). Effects of development and enculturation on number representation in the brain. *Nature Reviews. Neuroscience*, *9*(4), 278–291.

Ansari, D., Lyons, I. M., van Eimeren, L., & Xu, F. (2007). Linking visual attention and number processing in the brain: The role of the temporo-parietal junction in small and large symbolic and nonsymbolic number comparison. *Journal of Cognitive Neuroscience*, *19*(11), 1845–1853.

Brannon, E. (2006). The representation of numerical magnitude. *Current Opinion in Neurobiology*, *16*, 222–229.

Dehaene, S. (1997). *The number sense.* Oxford: 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*(2), 314–326.

Dehaene, S., & Cohen, L. (1995). Towards an anatomical and functional model of number processing. *Mathematical Cognition*(1), 83–120.

Dehaene, S., Dehaene-Lambertz, G., & Cohen, L. (1998). Abstract representations of numbers in the animal and human brain. *Trends in Neurosciences*, *21*(8), 355–361.

Gilmore, C., Attridge, N., & Inglis, M. (2011). Measuring the approximate number system. *Quarterly Journal of Experimental Psychology*, *64*(11), 2099–2109.

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

Holloway, I., & 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.

Hubbard, E. M., Diester, I., Cantlon, J. F., Ansari, D., Opstal, F., & Troiani, V. (2008). The evolution of numerical cognition: From number neurons to linguistic quantifiers. *The Journal of Neuroscience*, *28*(46), 11819.

Inglis, M., Attridge, N., Batchelor, E. S., & Gilmore, C. (2011). Non-verbal number acuity correlates with symbolic mathematics achievement: But only in children. *Psychonomic Bulletin & Review*, *18*(6), 1222–1229.

Landerl, K., Fussenegger, B., Moll, K., & Willburger, E. (2009). Dyslexia and dyscalculia: Two learning disorders with different cognitive profiles. *Journal of Experimental Child Psychology*, *103*(3), 309–324.

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.

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*(2), 154–161.

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

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

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature*, *215*(109), 1519–1520.

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.

Mussolin, C., Mejias, S., & Noel, M. P. (2010). Symbolic and nonsymbolic number comparison in children with and without dyscalculia. *Cognition*, *115*(1), 10–25.

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

Price, G. R., Holloway, I., Räsänen, P., Vesterinen, M., & Ansari, D. (2007). Impaired parietal magnitude processing in developmental dyscalculia. *Current Biology*, *17*(24), 1042–1043.

Sasanguie, D., Defever, E., Van den Bussche, E., & Reynvoet, B. (2010). The reliability of and the relation between non-symbolic numerical distance effects in comparison, same-different judgments and priming. *Acta Psychologica*, *136*(1), 73–80.

Straw, A. D. (2008). Vision egg: An open-source library for realtime visual stimulus generation. *Frontiers in Neuroinformatics*, doi:10.3389/neuro.11.004.2008.

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

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III Tests of Achievement.* Itasca, IL: Riverside Publishing.