Acta Psychologica 134 (2010) 154–161

Contents lists available at ScienceDirect

Acta Psychologica

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

Challenging the reliability and validity of cognitive measures: The case of the numerical distance effect

Erin A. Maloney [a,*], Evan F. Risko [b], Frank Preston [a], Daniel Ansari [c], Jonathan Fugelsang [a]

[a] Department of Psychology, University of Waterloo, Canada
[b] Department of Psychology, University of British Columbia, Canada
[c] Department of Psychology, University of Western Ontario, Canada

ARTICLE INFO

Article history:
Received 4 September 2009
Received in revised form 15 January 2010
Accepted 22 January 2010
Available online 24 February 2010

PsycINFO classification:
2340

Keywords:
Numerical Distance Effect
Numerical Cognition
Reliability
Numerical Processing

ABSTRACT

The numerical distance effect (NDE) is one of the most robust effects in the study of numerical cognition. However, the validity and reliability of distance effects across different formats and paradigms has not been assessed. Establishing whether the distance effect is both reliable and valid has important implications for the use of this paradigm to index the processing and representation of numerical magnitude in both behavioral and neuroimaging studies. In light of this, we examine the reliability and validity of frequently employed variants (and one new variant) of the numerical comparison task: two symbolic comparison variants and two nonsymbolic comparison variants. The results of two experiments demonstrate that measures of the NDE that use nonsymbolic stimuli are far more reliable than measures of the NDE that use symbolic stimuli. With respect to correlations between measures, we find evidence that the NDE that arises using symbolic stimuli is uncorrelated with the NDE that is elicited by using nonsymbolic stimuli. Results are discussed with respect to their implications for the use of the NDE as a metric of numerical processing and representation in research with both children and adults.

1. Introduction

The numerical distance effect (NDE) is obtained in tasks where participants are asked to perform relative magnitude judgements. In such experiments, participants are faster and more accurate at indicating which of the two numbers is larger when the numerical distance separating the two numbers is relatively large (e.g., 2 vs. 9), compared to when it is small (e.g., 8 vs. 6; Dehaene, Dupoux, & Mehler, 1990; Moyer & Landauer, 1967). Several models have been proposed to relate the NDE to numerical representation, for example, the "compressed number line" model (Dehaene, 1992), and the "numerosity code" model (Zorzi & Butterworth, 1999). Although the models differ in their characterization of the mental representation of quantity that is indexed by the numerical distance effect, they converge on the notion that the NDE provides an important metric for indexing the representation of numerical magnitude. This view of the NDE explains its frequent use to examine the processing and representation of numerical magnitude in both behavioral and neuroimaging studies.

There are multiple task variants used to elicit the NDE. For example, participants might be asked to compare two sets of nonsymbolic stimuli, such as squares (Holloway & Ansari, 2009), two Arabic digits (Ansari & Dhital, 2006; Ansari, Garcia, Lucas, Hamon, & Dhital, 2005; Dehaene, 1996), or one Arabic digit to a standard (i.e., the number 5; Dehaene, 1996; Libertus, Woldorff, & Brannon, 2007; Temple & Posner, 1998). Many researchers simply assume that because these variants produce the same pattern of data (i.e., a NDE), that each of these variants must be indexing the same stable underlying process. However, recent research has suggested that this assumption is invalid in at least one case of the NDE. Specifically, Holloway and Ansari (2009) presented a group of 6–8 year old children with two versions of the number comparison task, a symbolic and a nonsymbolic version. In the symbolic version of the task, Arabic digits were used. In the nonsymbolic version, arrays of squares were presented. Holloway and Ansari (2009) demonstrated that while symbolic stimuli (Arabic digits) and nonsymbolic stimuli (squares) both elicit an NDE, an individual's NDE on the symbolic version does not correlate with their NDE on the nonsymbolic version. This result raises important questions about the convergent validity of the NDE elicited by different stimulus and presentation formats (i.e., the extent to which measures of the NDE are in fact measuring the same underlying construct). In the present investigation we sought to investigate the degree to which multiple measures of the NDE are indexing the same underlying processes. To accomplish this, we used multiple variants of the numerical comparison task, each known to produce an NDE, and tested whether the size of a participant's NDE on

* Corresponding author. Tel.: +1 519 277 1428.
E-mail address: eamalone@uwaterloo.ca (E.A. Maloney).

0001-6918/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.actpsy.2010.01.006

one variant of the task correlated with the size of their NDE on the other variants of the task.

Another issue that arises when considering the validity of measures is whether or not they are reliable. Reliability is a fundamental psychometric property that should be determined in the measurement of any theoretically important empirical construct. Certainly, when researchers are developing scales, reliability is of the utmost importance. The issue of reliability of measures used in mainstream cognitive psychology, however, is typically neglected. One potential reason for this in experimental cognitive psychology is the implicit assumption that cognitive processes are inherently reliable because they typically produce robust and replicable empirical phenomena. This assumption may also, in large part, reflect the widespread belief that many cognitive processes have automatic components that are expected to unfold in a stable and consistent manner (Stolz, Besner, & Carr, 2005).

Recently, researchers have begun to examine the reliability of various cognitive measures with some surprising results (e.g., Borgmann, Risko, Stolz, & Besner, 2007; Stolz et al., 2005; Waechter, Stolz, & Besner, in press). This recent work has relied on a test–retest assessment of reliability correlating the size of a given effect for the first half and second half of the experiment. If a measure is reliable, then scores at time one (e.g., the first half) will be highly predictive of scores at time two (e.g., the second half) of that test. Using this approach, Stolz et al. (2005) found semantic priming to be generally unreliable, whereas Waechter, Stolz, & Besner (in press) found repetition priming to be comparatively reliable. In research investigating the Simon effect (Simon, 1990), Borgmann et al. (2007) identified a context where the effect proved reliable (when compatibility proportion is high) and a context where it was unreliable (when compatibility proportion is low). Taken together, these findings reveal that the implicit assumption that cognitive processes are inherently reliable is not always supported by the data.

While assessing reliability is important in general, for the present purposes it is particularly important because when an effect is not reliable it is limited in how strongly it can correlate with other measures. The fact that many measures in cognitive psychology are used as corollaries in individual difference studies makes it important to know the reliability of these measures. However, if a measure is unreliable, the likelihood of detecting between group differences when using this particular paradigm is rather low even if those differences do exist (Kopriva & Shaw, 1991). As such, it makes the interpretation of a null result rather difficult when one does not know how reliable the measure used is. Thus assessing whether measures of cognitive processes, such as numerical cognition, are reliable and valid is not only of methodological importance but also has a significant bearing on the theoretical explanations and models derived from data obtained with such dependent variables.

In two experiments, we assess the convergent validity of the NDE by comparing participant's NDE on multiple variants of the number comparison tasks (three variants in Experiment 1 and four variants in Experiment 2). Further, we assess the reliability of each version by employing a test retest approach where we compare participant's NDE in the first half of the experiment to their NDE in the second half of the experiment.

2. Experiment 1

2.1. Methods

2.1.1. Participants

Forty-eight undergraduate students from the University of Waterloo participated and were either granted experimental credit towards a course or were paid $6.

2.1.2. Stimuli, apparatus and procedure

The data were collected on a Pentium 4 PC computer running E-Prime 1.1 (Schneider, Eschman, & Zuccolotto, 2001). Stimuli were displayed on a 19 in. monitor. Participants participated in three separate tasks (lower/higher than 5, symbolic comparison, and nonsymbolic comparison) that were presented in a counterbalanced order across participants. The three tasks are described in detail below.

2.1.2.1. Lower/higher than five (L/H5). Each trial began with a fixation point that remained on the screen for 500 ms. A display containing a single Arabic digit at fixation was presented. Numbers ranged from 1 to 4 and from 6 to 9. Participants were told to identify whether the presented number was lower than five or higher than five by pressing the "A" key to denote lower and the "L" key to denote higher. The number remained on the screen until the participants made a button press. There were two blocks each with 80 stimulus displays making for a total of 160 trials. The numerical distance between the stimuli and the number 5 ranged from 1 to 4, with 40 comparison trials total per distance. Trial composition was identical for blocks 1 and 2 and stimulus displays were presented randomly within each block.

2.1.2.2. Symbolic comparison. Each trial began with a fixation point that remained on the screen for 500 ms. A display containing two Arabic digits was presented. Numbers ranged from 1 to 4 and from 6 to 9. Participants were told to identify which of the two numbers was numerically larger by pressing the "A" key to denote that the number on the left was larger and the "L" key to denote that the number on the right was larger. There were two blocks each with 80 stimulus displays making for a total of 160 trials. The numerical distance between the stimuli ranged from 1 to 4, with 40 comparison trials total per distance. As with the L/H5 task, trial composition was identical for blocks 1 and 2 and stimulus displays were presented randomly within each block.

2.1.2.3. Nonsymbolic comparison. Each trial began with a fixation point that remained on the screen for 500 ms. A display containing two white square boxes was presented. In each white box there were a number of black squares. The number of black squares ranged from 1 to 4 and from 6 to 9. Participants were told to identify which of the two white square boxes contained more black squares by pressing the "A" key to denote that the box on the left contained more squares and the "L" key to denote that the box on the right contained more squares. There were two blocks each with 80 stimulus displays making for a total of 160 trials. The numerical distance between the stimuli ranged from 1 to 4, with 40 comparison trials total per distance. Stimulus displays were presented randomly within each block and contained the same numerical value pairs as the symbolic comparison tasks. Furthermore, the individual area, total area, and density of the squares were varied to ensure that participants could not reliably use non-numerical cues to make a correct decision (see Holloway & Ansari, 2009 for details on the stimuli and previous usage).

2.2. Results

RTs and errors were analyzed across participants with block and numerical distance as within-subject factors. Trials on which there was an incorrect response were removed prior to RT analysis (3.3% in the L/H5, 2.8% in the symbolic, and 5.0% in the nonsymbolic variants, respectively). The remaining RTs were submitted to a data trimming procedure that uses a 2.5 standard deviation cut-off in each cell for each subject. The trimming procedure is run recursively until there are no longer any cases that lay more than 2.5 standard deviations from the mean in any cell (Van Selst &

Table 1
Mean RTs (ms) and range by block at distances of 1 and 4 for Experiment 1.

Variant	Block 1				Block 2			
	1	Range	4	Range	1	Range	4	Range
L/H5	536	392.7–806.0	484	362.2–845.8	535	411.7–729.0	471	371.7–621.3
Symbolic	543	393.0–856.0	480	379.0–676.5	520	371.7–738.5	469	363.9–621.5
Nonsymbolic	1301	442.9–3088.1	710	417.0–1690.8	1132	417.4–2678.3	609	357.4–1121.9

Jolicœur, 1994). This resulted in the removal of an additional 2.7% of the data in the L/H5, 3.4% in the symbolic, and 2.6% in the nonsymbolic variants, respectively. In addition, given that correlational analyses are heavily influenced by extreme scores, we first calculated distance effect scores (RT at a distance of 1 minus RT at a distance of 4) for each participant in each variant of the task. We then removed any participants with a distance effect that fell 4 or more standard deviations above or below the mean in either variant. This resulted in the removal of one participant whose distance effect in the L/H5 task fell 4.7 standard deviations above the mean.

For each analysis we present the RT data followed by the error data. We report the error data for completeness, however, it is important to note that little can be made of the error data due to the fact that very few errors were made and the size of the NDE in errors is small. Critically, studies of the NDE typically focus on RT and not errors.

2.2.1. Numerical distance effect analysis

Next, we determined that each variant of the task did, in fact, elicit the NDE. These data are reported below. See Tables 1 and 2 for mean distance effects in RTs and accuracy, respectively.

2.2.1.1. Lower/higher than five (L/H5). A 2 (block: 1 vs. 2) × 2 (distance: 1 vs. 4)[1] ANOVA conducted on the RT data yielded no main effect of block, $F(1, 47) = .80$, $MSE = 3227.5$, $p > 0.5$, a main effect of distance, $F(1, 47) = 99.6$, $MSE = 826.4$, $p < .001$, and no block × distance interaction, $F(1, 47) = 1.0$, $MSE = 1475.9$, $p > .05$. A parallel ANOVA conducted on the error data yielded no main effect of block, $F(1, 47) = 2.6$, $MSE = 17.5$, $p > .05$, a main effect of distance, $F(1, 47) = 35.1$, $MSE = 26.2$, $p < .001$, and no block × distance interaction, $F(1, 47) = 1.7$, $MSE = 15.4$, $p > .05$.

2.2.1.2. Symbolic comparison. A 2 (block: 1 vs. 2) × 2 (distance: 1 vs. 4) ANOVA conducted on the RT data yielded a main effect of block, $F(1, 47) = 6.1$, $MSE = 2298.4$, $p = .02$, a main effect of distance, $F(1, 47) = 96.8$, $MSE = 1599.0$, $p < .001$, and no block × distance interaction, $F(1, 47) = 2.1$, $MSE = 764.8$ $p > .05$. A parallel ANOVA conducted on the error data yielded no main effect of block, $F(1, 47) = .08$, $MSE = 13.6$ $p > .05$, a main effect of distance, $F(1, 47) = 37.0$, $MSE = 21.9$, $p < .001$, and no block × distance interaction, $F(1, 47) = 3.0$, $MSE = 12.4$ $p = .09$.

2.2.1.3. Nonsymbolic comparison. A 2 (block: 1 vs. 2) × 2 (distance: 1 vs. 4) ANOVA conducted on the RT data yielded a main effect of block, $F(1, 47) = 16.2$, $MSE = 5698.9$, $p < .01$, a main effect of distance, $F(1, 47) = 46.9$, $MSE = 317473.7$, $p < .001$, and no block × distance interaction $F(1, 47) = 2.5$, $MSE = 21925.9$, $p > .05$. A parallel ANOVA conducted on the error data yielded no main effect of block, $F(1, 47) = .66$, $MSE = 28.5$, $p > .05$, a main effect of distance, $F(1, 47) = 7.8$, $MSE = 68.1$, $p < .001$, and no block × distance interaction $F(1, 47) = .12$, $MSE = 16.7$, $p > .05$.

[1] Only distances 1 and 4 are used in the ANOVA because distance effects are calculated as RTs (or errors) at a distance of 1 minus RTs (or errors) at a distance of 4 for subsequent analyses. The pattern of data did not change when the ANOVA was calculated using all four distances, nor did the pattern change when calculated using the individual slopes that relate numerical distance to RTs and errors. The pattern of data was also not altered when we calculated NDEs as RTs (or errors) at a distance of 1 minus RTs (or errors) at a distance of 4 divided by average RT. This is true in both the reliability and validity calculations and for both Experiments 1 and 2. Furthermore, calculating the effect as 1–4 gives us the largest range and thus the highest likelihood of detecting correlations.

Table 2
Mean accuracy (% correct) and range by block at distances of 1 and 4 for Experiment 1.

Variant	Block 1				Block 2			
	1	Range	4	Range	1	Range	4	Range
L/H5	94	80–100	98	90–100	93	70–100	98	90–100
Symbolic	95	80–100	100	95–100	95	85–100	98	90–100
Nonsymbolic	91	60–100	94	60–100	92	65–100	95	45–100

2.2.2. Convergent validity

A mean NDE was calculated for each participant by subtracting the mean RT (or accuracy) for a distance of 4 from the mean RT (or accuracy) for a distance of 1. We collapsed across block when calculating mean NDE scores for the convergent validity analysis. See Tables 3 and 4 for between variant correlations in RTs and accuracy, respectively.

2.2.2.1. Symbolic comparison variants. The correlation between each participant's mean NDE score on the L/H5 variant and their mean NDE score on the symbolic comparison variant was only found to be significant in errors, $r(46) = .49$, $p < .01$.

2.2.2.2. Symbolic vs. nonsymbolic. A correlation was conducted between each participant's mean NDE score on the L/H5 variant, their mean NDE score on the symbolic comparison variant, and their mean NDE score on the nonsymbolic comparison variant. None of the correlations between the symbolic and nonsymbolic variants were significant. This was found to be true for both RT and errors.

2.2.3. Reliability

A NDE was calculated for each block by subtracting the mean RT (or accuracy) for a distance of 4 from the mean RT (or accuracy) for a distance of 1. A significant correlation between Blocks 1 and 2 scores indicates reliability (e.g., Stolz et al., 2005). See Table 5 for mean reliability effects in RTs and accuracy.

2.2.3.1. Lower/higher than five (L/H5). The correlation between a participant's NDE in Block 1 and their NDE in Block 2 was not significant in RTs, $r(46) = .06$, $p > .05$, and marginally significant in errors, $r(46) = .27$, $p = .06$.

2.2.3.2. Symbolic comparison. The correlation between a participant's NDE in Block 1 and their NDE in Block 2 was found to be significant in RTs, $r(46) = .38$, $p < .01$, and marginally significant in errors, $r(46) = .28$, $p = .05$.

2.2.3.3. Nonsymbolic comparison. The correlation between a participant's NDE on Block 1 and their NDE on Block 2 was found to be

Table 3
Between variant Correlation (ms) for Experiment 1.

Block 1	Block 2		
	L/H5	Symbolic	Nonsymbolic
L/H5	1.00	0.19	−0.12
Symbolic		1.00	0.05
Nonsymbolic			1.00

Table 4
Between variant correlations (% correct) for Experiment 1.

Block 1	Block 2		
	L/H5	Symbolic	Nonsymbolic
L/H5	1.00	0.49*	−0.11
Symbolic		1.00	0.15
Nonsymbolic			1.00

*Denotes significant at the $p < .05$ level.

Table 5
Within variant correlations for RTs (ms) and accuracy (% accurate) for Experiment 1.

Variant	Reliability	
	RTs	% Accuracy
L/H5	0.06	0.27+
Symbolic	0.38*	0.28+
Nonsymbolic	0.88*	0.61*

*Denotes significant at the $p < .05$ level.
+Denotes marginally significant.

significant in RTs, $r\ (46) = .89$, $p < .01$, and errors, $r\ (46) = .61$, $p < .01$.

2.3. Summary

In Experiment 1 we assessed the degree to which the numerical distance effect (NDE) elicited by one variant of the numerical comparison task correlates with performance on other versions of this task (i.e., the convergent validity of the NDE). In addition, we assessed the degree to which a participants' NDE on the first half of a variant correlated with their NDE on the second half of that variant (i.e., the reliability of the NDE). We did not find a significant correlation between the two symbolic variants (those which used Arabic digits) in RTs but we did in accuracy. We also did not find a correlation in RTs or accuracy between the symbolic and nonsymbolic comparison variants. In terms of reliability, the L/H5 variant was unreliable in RTs and only marginally reliable in errors. The symbolic comparison variant is statistically reliable, but in terms of a psychometric battery the low correlations would fall far short of useful (Murphy & Davidshofer, 2005). The nonsymbolic comparison variant, however, is very reliable in both RT and errors.

Arguably, the L/H5 paradigm, where an Arabic digit is compared to a fixed standard, can be considered more different from the two comparison paradigms, than the two comparison paradigms are from each other. For instance, in the L/H5 paradigm participants are comparing a stimulus to a fixed standard, rather than to a comparable stimulus. Furthermore, in the L/H5 paradigm the participant has to differentiate between small and large, whereas in the other paradigms only the larger number had to be indicated. In order to best equate the symbolic and nonsymbolic task variants and assess the validity and reliability of the NDE, we therefore ran a second experiment where we administered a new nonsymbolic variant of the L/H5 task. Thus, Experiment 2 served both as a replication of Experiment 1 and as an extension by better equating the symbolic and nonsymbolic task variants.

3. Experiment 2

3.1. Methods

3.1.1. Participants

Forty-eight undergraduate students from the University of Waterloo participated and were granted experimental credit towards a course.

3.1.2. Stimuli, apparatus and procedure

The stimuli, apparatus, and procedure used in Experiment 2 were identical to those used in Experiment 1 with one exception. In Experiment 2, we added a fourth task, nonsymbolic lower/higher than 5. In this task each trial began with a fixation point that remained on the screen for 500 ms. A display containing a number of black squares was presented. The number of black squares ranged from 1 to 4 and from 6 to 9. Participants were told to identify whether the number of boxes presented was lower than five or higher than five by pressing the "A" key to denote lower and the "L" key to denote higher. The display remained on the screen until the participants made a button press. There were two blocks each with 80 stimulus displays making for a total of 160 trials. The numerical distance between the stimuli and the number 5 ranged from 1 to 4, with 40 comparison trials total per distance. Stimulus displays were presented randomly within each block. Furthermore, the individual area, total area, and density of the squares were varied to ensure that participants could not reliably use non-numerical cues to make a correct decision. For simplicity, the present variant will be referred to as nonsymbolic lower/higher than 5. The version of this variant which was also tested in Experiment 1 will henceforth be referred to as symbolic lower/higher than 5.

3.2. Results

RTs and errors were analyzed across participants, with numerical distance and task variant as within-subject factors. Trials on which there was an incorrect response were removed prior to RT data analysis (2.1% in symbolic L/H5, 3.8% in nonsymbolic L/H5, 4.0% in symbolic comparison, and 4.5% in nonsymbolic comparison). The remaining RTs were submitted to a recursive data-trimming procedure using a 2.5 standard deviation cut-off in each cell resulting in the removal of an additional 2.9% of trials in symbolic L/H5, 3.1% in nonsymbolic L/H5, 3.8% in symbolic comparison, and 5.5% in nonsymbolic comparison, respectively. In addition, given that correlational analyses are heavily influenced by extreme scores, we first calculated distance effect scores (RT at a distance of 1 minus RT at a distance of 4) for each participant in each variant of the task. As per Experiment 1, we would have then removed any participants with a distance effect that fell 4 or more standard deviations above or below the mean in either variant. However, there were no such participants.

3.2.1. Numerical distance effect analysis

Next, we determined that each variant of the task did, in fact, elicit an NDE. These data are reported below. See Tables 6 and 7 for mean distance effects in RTs and accuracy, respectively.

3.2.1.1. Symbolic lower/higher than five (L/H5). A 2 (block: 1 vs. 2) × 2 (distance: 1 vs. 4) ANOVA conducted on the RT data yielded a main effect of block, $F(1, 47) = 5.4$, $MSE = 3437.1$, $p < .05$, a main effect of distance, $F(1, 47) = 69.5$, $MSE = 2590.0$, $p < .001$, and no block × distance interaction, $F(1, 47) = 1.9$, $MSE = 2119.2$, $p > .05$. A parallel ANOVA conducted on the error data yielded a marginal effect of block, $F(1, 47) = 3.8$, $MSE = 39.6$, $p = .06$, a main effect of

distance, $F(1, 50) = 46.2$, $MSE = 41.0$, $p < .01$, and no block × distance interaction, $F < 1$.

3.2.1.2. Nonsymbolic lower/higher than five (L/H5). A 2 (block: 1 vs. 2) × 2 (distance: 1 vs. 4) ANOVA conducted on the RT data yielded a main effect of block, $F(1, 47) = 25.4$, $MSE = 8564.2$, $p < .01$, a main effect of distance, $F(1, 47) = 75.4$, $MSE = 1286.2$, $p < .01$, and a block × distance interaction, $F(1, 47) = 21.4$, $MSE = 3445.6$, $p < .01$ in which the NDE was found to be smaller in Block 2 than in Block 1. A parallel ANOVA conducted on the error data yielded no main effect of block, $F(1, 47) = 1.8$, $MSE = 28.4$, $p > .05$, a main effect of distance, $F(1, 47) = 51.9$, $MSE = 33.7$, $p < .01$, and no block × distance interaction, $F(1, 47) = 2.1$, $MSE = 20.1$, $p > .05$.

3.2.1.3. Symbolic comparison. A 2 (block: 1 vs. 2) × 2 (distance: 1 vs. 4) ANOVA conducted on the RT data yielded no main effect of block, $F(1, 47) = 1.1$, $MSE = 3247.9$, $p > .05$, a main effect of distance, $F(1, 47) = 118.9$, $MSE = 1323.5$, $p < .01$, and no block × distance interaction, $F(1, 47) < 1$. A parallel ANOVA conducted on the error data yielded a marginal effect of block, $F(1, 47) = 3.8$, $MSE = 39.6$, $p = .06$, a main effect of distance, $F(1, 47) = 46.9$, $MSE = 36.1$, $p < .01$, and no block × distance interaction, $F < 1$.

3.2.1.4. Nonsymbolic comparison. A 2 (block: 1 vs. 2) × 2 (distance: 1 vs. 4) ANOVA conducted on the RT data yielded a main effect of block, $F(1, 47) = 22.9$, $MSE = 52298.5$, a main effect of distance, $F(1, 47) = 24.9$, $MSE = 425814.7$, and a block × distance interaction, $F(1,47) = 7.0$, $MSE = 39319.4$, $p < .05$ in which the NDE was found to be smaller in Block 2 than in Block 1. A parallel ANOVA conducted on the error data yielded no main effect of block, $F(1, 47) = 1.5$, $MSE = 16.7$, $p > .05$, a main effect of distance, $F(1, 47) = 79.5$, $MSE = 4.3$, and no block × distance interaction, $F < 1$.

3.2.2. Convergent validity

A mean NDE was calculated for each participant by subtracting the mean RT (or accuracy) for a distance of 4 from the mean RT (or accuracy) for a distance of 1. We collapsed across block when calculating mean NDE scores for the validity analysis. See Tables 8 and 9 for validity effects in RTs and accuracy, respectively.

3.2.2.1. Symbolic comparison variants. A correlation was conducted between each participant's mean NDE score on the symbolic L/H5 variant and their mean NDE score on the symbolic comparison variant. The correlation was found to be only marginally significant, $r(46) = .25$, $p = .08$. A parallel analysis conducted on the error data was found to be significant, $r(46) = -.29$, $p < .05$, however in the negative direction.

Table 8
Between variant correlation (ms) for Experiment 2.

Block 1	Block 2			
	Symbolic L/H5	Nonsymbolic L/H5	Symbolic comparison	Nonsymbolic comparison
Symbolic L/H5	1.00	0.04	0.25[+]	0.21
Nonsymbolic L/H5		1.00	0.18	0.56[*]
Symbolic comparison			1.00	0.05
Nonsymbolic comparison				1.00

[*] Denotes significant at the $p < .05$ level.
[+] Denotes marginally significant.

3.2.2.2. Nonsymbolic comparison variants. A correlation was conducted between each participant's mean NDE score on the nonsymbolic L/H5 variant and their mean NDE score on the nonsymbolic comparison variant. The correlation was significant in RTs, $r(46) = .56$, $p < .01$, and marginally significant in errors, $r(46) = .29$, $p = .05$.

3.2.2.3. Symbolic vs. nonsymbolic. A correlation was conducted between each participant's mean NDE score on the symbolic L/H5 variant, the nonsymbolic L/H5 variant, the symbolic comparison variant, and the nonsymbolic comparison variant. None of the correlations between the symbolic and nonsymbolic variants were significant in RTs (as can be seen in Table 8, the largest correlation, that between symbolic L/H5 and nonsymbolic comparison, was $r = .21$). A parallel analysis conducted on the error data yielded only one significant correlation: the correlation between the symbolic L/H5 variant and the nonsymbolic comparison variant, $r(46) = .31$, $p < .05$.

3.2.3. Reliability

3.2.3.1. Symbolic lower/higher than five (L/H5). The correlation between a participant's NDE in Block 1 and their NDE in Block 2 was not significant in RTs, $r(46) = .10$, $p > .05$ and marginally significant in errors, $r(46) = .27$, $p = .09$.

Table 6
Mean RTs (ms) and range by block at distances of 1 and 4 for Experiment 2.

Variant	Block 1				Block 2			
	1	Range	4	Range	1	Range	4	Range
Symbolic L/H5	549	518–580	479	457–500	520	486–555	468	449–488
Nonsymbolic L/H5	690	628–752	506	480–526	581	544–618	481	459–503
Symbolic comparison	545	511–579	484	455–512	533	497–569	479	449–509
Nonsymbolic comparison	1179	924–1435	633	570–696	946	728–1164	551	502–600

Table 7
Mean accuracy (% accuracy) and range by block at distances 1 and 4 for Experiment 2.

Variant	Block 1				Block 2			
	1	Range	4	Range	1	Range	4	Range
Symbolic L/H5	92	89–94	98	96–100	90	86–94	96	94–98
Nonsymbolic L/H5	91	88–94	96	95–98	89	87–91	96	94–98
Symbolic comparison	92	89–95	96	93–100	93	90–95	96	93–98
Nonsymbolic comparison	90	87–93	98	96–100	89	86–92	98	96–100

Table 9
Between variant correlations (% accurate) for Experiment 2.

Block 1	Block 2			
	Symbolic L/H5	Nonsymbolic L/H5	Symbolic comparison	Nonsymbolic comparison
Symbolic L/H5	1.00	0.15	−0.29*	0.31*
Nonsymbolic L/H5		1.00	0.16	0.28+
Symbolic comparison			1.00	0.18
Nonsymbolic comparison				1.00

*Denotes significant at the $p < .05$ level.
+Denotes marginally significant.

Table 10
Within variant correlations for RTs (ms) and accuracy (% accurate).

Variant	Reliability	
	RTs	% Accurate
Symbolic L/H5	0.10	0.25+
Nonsymbolic L/H5	0.64*	0.54*
Symbolic comparison	0.14	0.00
Nonsymbolic comparison	0.84*	0.54*

*Denotes significant at the $p < .05$ level.
+Denotes marginally significant.

3.2.3.2. Nonsymbolic lower/higher than five (L/H5). The correlation between a participant's NDE in Block 1 and their NDE in Block 2 was highly significant in RTs, $r(46) = .84$, $p < .01$, and marginal in errors, $r(46) = .26$, $p = .08$.

3.2.3.3. Symbolic comparison. The correlation between a participant's NDE on Block 1 and their NDE on Block 2 was not significant in RTs, $r(46) = .14$, $p > .05$ or errors, $r(46) = .01$, $p > .05$.

3.2.3.4. Nonsymbolic comparison. The correlation between a participant's NDE on Block 1 and their NDE on Block 2 was highly significant in RTs, $r(46) = .84$, $p < .01$ and errors, $r(46) = .54$, $p < .01$ (Table 10).

3.3. Summary

Experiment 2 replicated the main results of Experiment 1 and extended them by assessing the reliability of a new variant (nonsymbolic L/H5) of the numerical comparison task. With respect to RTs, we found that the two symbolic tasks (those which used Arabic digits) correlate only marginally. However, we found that performance on the two nonsymbolic variants significantly correlated with each other.

In terms of reliability, the new nonsymbolic L/H5 variant was significantly reliable on RTs. In addition, we replicated the findings of Experiment 1 concerning the reliability of the symbolic L/H5 task and the nonsymbolic comparison task. However, the symbolic comparison task, which was reliable in Experiment 1, was not reliable in Experiment 2.

4. Combined analysis of Experiments 1 and 2

Given there were a few discrepancies between the results from Experiments 1 and 2 (i.e., the reliability of the symbolic comparison task), we conducted analyses collapsing across the two data sets for the three common variant types (symbolic L/H5, symbolic comparison, and nonsymbolic comparison).

4.1. Convergent validity

4.1.1. Symbolic comparison variants

The correlation between each participant's mean NDE score on the symbolic L/H5 variant and their mean NDE score on the symbolic comparison variant was found to be significant in errors, $r(94) = .22$, $p < .05$, but not in RTs.

4.1.1.1. Symbolic comparison vs. nonsymbolic comparison. As is evident in Tables 11 and 12, neither of the symbolic variants and the nonsymbolic variants correlated with each other. This was true for both RTs and errors.

4.1.2. Reliability

4.1.2.1. Symbolic lower/higher than five (L/H5). The correlation between a participant's NDE on Block 1 and their NDE on Block 2 was not found to be statistically significant in RTs, $r(94) = .07$, $p > .05$, but was found to be significant in errors, $r(94) = .27$, $p < .01$.

4.1.2.2. Symbolic comparison. The correlation between a participant's NDE on Block 1 and their NDE on Block 2 was found to be statistically significant in RTs, $r(94) = .25$, $p < .05$, but not in errors, $r(94) = .10$, $p > .05$.

4.1.2.3. Nonsymbolic comparison. The correlation between a participant's NDE on Block 1 and their NDE on Block 2 was found to be statistically significant, $r(94) = .85$, $p < .01$. A parallel analysis conducted on the error data yielded a significant correlation, $r(94) = .61$, $p < .01$ (Table 13).

Table 11
Between variant correlations (ms) for combined data.

Block 1	Block 2		
	L/H5	Symbolic	Nonsymbolic
L/H5	1.00	0.22*	0.08
Symbolic		1.00	0.05
Nonsymbolic			1.00

+Denotes marginally significant.
*Denotes significant at the $p < .05$ level.

Table 12
Between variant correlations (% accuracy) for combined data.

Block 1	Block 2		
	L/H5	Symbolic	Nonsymbolic
L/H5	1.00	0.02	0.12
Symbolic		1.00	0.11
Nonsymbolic			1.00

Table 13
Within variant correlations (% accuracy) for combined data.

Variant	Reliability	
	RTs	% Accuracy
L/H5	0.07	0.27*
Symbolic	0.25*	0.10
Nonsymbolic	0.85*	0.61*

*Denotes significant at the $p < .05$ level.

4.2. Summary

The results from the combined analysis confirm that the NDEs elicited on the symbolic and nonsymbolic comparison variants, in fact, do not correlate with one another. In addition, the combined analysis further confirms that, in terms of split-half reliability, the symbolic L/H5 variant is unreliable while the symbolic comparison variant and the nonsymbolic comparison variant are reliable.

5. General discussion

The numerical distance effect (NDE) represents a frequently used measure in the field of numerical cognition and plays a critical role in many theories of magnitude representation (see Ansari, 2008; Dehaene, 1997; Nieder, 2005 for reviews). While Holloway and Ansari (2009) assessed the convergent validity of different numerical comparison variants, the reliability of NDEs across different formats and paradigms has not been assessed. Establishing both the validity and the reliability of NDEs has important implications for how we interpret the results in both behavioral and neuroimaging studies. Against this background, we investigated whether or not the NDE elicited using symbolic stimuli and the NDE elicited using nonsymbolic arise as a function of the same or different underlying mechanisms. Results from two studies suggest that the mechanisms giving rise to the symbolic NDE and the nonsymbolic NDE are not the same, since the two distance effects were consistently not found to correlate with one another. The present study further reveals that, in terms of split-half reliability, the symbolic L/H5 variant has low reliability in RTs but is reliable in errors while the nonsymbolic L/H5 variant is reliable both in RTs and errors. Furthermore, both the symbolic comparison variant and the nonsymbolic comparison variant are reliable.

The observation that the NDEs that arise when using symbolic stimuli and the NDEs that arise when using nonsymbolic stimuli do not correlate in RTs is consistent with Holloway and Ansari's (2009) data from primary school children. We are thus able to extend their data to a sample of adults. Furthermore, the present investigation allowed us to assess whether or not the lack of correlation reported in Holloway and Ansari (2009) was due to a lack of reliability inherent in either variant. They assumed that the lack of a correlation between a participant's NDE in symbolic numerical comparison and their NDE in nonsymbolic numerical comparison meant that the two NDEs arose due to different underlying mechanisms (i.e., they are not the same NDE). However, they did not rule out the possibility that even though the two measures were both indexing the same NDE, one or both were unreliable. In addition to addressing the reliability of the NDE, we extend the findings of Holloway and Ansari (2009) by asserting that the lack of correlation between the symbolic and the nonsymbolic comparison variants is likely *not* due to a lack of reliability of either variant used. Thus, it appears to be the case that the NDE observed with symbolic variants and the NDE observed with nonsymbolic variants of the task are arising as a result of different underlying mechanisms.

5.1. Implications for individual and group differences

The results of the present investigation have important implications for researchers who use the NDE to study individual and group differences. Specifically, using an unreliable measure makes it more difficult to detect between group differences in mean NDE, between group differences in NDE-related brain activation, and between group differences in NDE-related measures on diagnostic batteries. Given that using measures with low reliability can substantially diminish the likelihood of detecting existing differences between groups, researchers should be cautious in interpreting any null effects when using symbolic numerical comparison tasks. We caution against the symbolic comparison variant because, while it is statistically reliable, a split-half correlation of .25 is still considered to be very poor (Murphy & Davidshofer, 2005). While a split-half correlation of .25 is still reliable enough that it can correlate with other measures (certainly in our combined analysis the symbolic comparison variant and the symbolic L/H5 variant did correlate), the power to detect a correlation between two measures diminishes with decreasing reliability. Our results indicate that we should exercise caution when trying to infer results from the various tasks that produce an NDE.

5.2. Theoretical implications

Beyond the methodological implications of the present findings, the data reported in the present paper have a number of potential theoretical implications. The findings suggest that symbolic and nonsymbolic distance effects not only differ in terms of their reliability, but are also uncorrelated with one another. This implies that these effects index different cognitive processes during numerical magnitude comparisons that vary as a function of stimulus format. It has been proposed that while symbolic and nonsymbolic numerical magnitudes rely on a common internal, place-coded (each number occupies a specific place on a 'mental number line') representation, the input-to-representation mapping pathways differ between symbolic and nonsymbolic formats for the representation of numerical magnitude. Specifically Verguts and Fias (2004) demonstrated, using computational modeling, that nonsymbolic numerical magnitude processing involves an intermediate step of generating a summation code, whereby nonsymbolic inputs are summed so that they can be linearly transformed into an internal, format-independent, place-coded representation. Such a step is not computationally necessary for symbolic representations. Thus, symbolic inputs (such as Arabic numerals) can be directly mapped onto an internal-place-coded representation. Thus, both symbolic and nonsymbolic stimuli will both activate the same place-coded magnitude representation on the number line but the input-to-representation pathways are different. The notion of different input-to-representation pathways for symbolic and nonsymbolic representation of numerical magnitude has been substantiated by a number of recent functional brain imaging studies (Holloway, Price, & Ansari, 2010; Santens, Roggerman, Fias, & Verguts, 2010). Furthermore, the lack of an intermediate step for summation coding also implies that the tuning curves of the place-coded symbolic representations are narrower than the nonsymbolic tuning curves. Therefore, symbolic, place-coded internal representation of numerical magnitude are hypothesized to be characterized by less overlap between adjacent place-coded representations on the internal 'mental number line' than nonsymbolic, place-coded representations. In essence, the Arabic digit "2" and two squares will both activate the same magnitude "two" on the internal number line. However, because we can go directly from the Arabic digit "2" to the corresponding magnitude and do not need the intermediate step of summation coding that is required when we process nonsymbolic stimuli, the tuning curve around that representation is more precise than when it is accessed using two squares. Taken together, this account predicts that both symbolic and nonsymbolic representations are place-coded, but that the tuning curves of numerical magnitudes on the place-coded representation are narrower for symbolic compared to symbolic representations. The present data are certainly consistent with this theory as, in Experiment 2, the mean size of the nonsymbolic NDE (306 ms is significantly larger than the mean size of the symbolic NDE (59 ms). This same pattern holds true for Experiment 1, as

well as in the combined analysis. Indeed, recent neuroimaging findings have indicated that the tuning curves for symbolic numerical magnitude in the left intraparietal sulcus are more precise than those for nonsymbolic numerical magnitude (Piazza, Izard, Pinel, Le Bihan, and Dehaene, 2004). It is thus possible that the lack of a correlation between symbolic and nonsymbolic distance effects in the present paper and in the earlier reported findings by Holloway and Ansari (2009) reflect differences in the way in which numerical magnitudes are accessed from symbolic and nonsymbolic numerical stimuli.

An alternative explanation for the lack of a correlation between symbolic and nonsymbolic magnitudes is that these index qualitatively different representations of numerical magnitude in the brain (for a discussion of stimulus-dependent, non-abstract representations of numerical magnitude see Cohen Kadosh and Walsh, 2007). So rather than the input-to-representation pathways differing as a function of the external format within which numerical magnitudes are represented, this theory predicts that each format is represented differently in the brain.

It is important to note that the extent to which the NDE indexes representations of numerical magnitude has been disputed. Specifically, Van Opstal et al. (2008) demonstrated that while a comparison distance effect can be found for both letters and numbers, a priming distance effect is only observed for numbers but not for letters. Against the background of these findings, Van Opstal et al. argue that the comparison distance effect does not index overlapping representations of numerical magnitude, since these cannot be assumed during the comparison of letters. Instead, these authors contend that the comparison distance effect (such as the one measured in both experiments above) indexes processes related to the resolution of the response alternatives during number comparison. It is thus possible that symbolic and nonsymbolic numerical magnitude comparison tasks do not index either different input-to-representation mapping pathways or qualitatively different representations, but instead reflect differences in the demands they place on the response–selection component of numerical magnitude comparison, which may give rise to a different distance effect. In future studies, these alternative explanations should be explored further by investigating the reliability and validity of symbolic and nonsymbolic priming distance effects as well as the exploration of these effects using functional neuroimaging.

6. Conclusion

We have assessed the convergent validity of the NDE and determined that although the NDE can be elicited using both symbolic and nonsymbolic stimuli, these NDEs do not correlate. We have also assessed the reliability of four different measures of the NDE. We demonstrated that the nonsymbolic variants are highly reliable, the symbolic numerical comparison variant is reliable (however, the Block 1–Block 2 correlations were low), and the symbolic L/H5 variant is unreliable. Taken together, these findings suggest that a great deal of caution should be exercised when using symbolic comparison tasks, in particular the L/H5 variant, to glean insight from null effects. Outside of the scope of numerical cognition, these findings highlight the importance of moving beyond solely looking for replicable measures towards an assessment of replication, reliability and construct validity of effects used to garner insights into cognitive processes.

Acknowledgments

This research was supported by Discovery Grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) to JF and DA, and both a NSERC and a Killam postdoctoral fellowship to EFR.

References

Ansari, D. (2008). Effect of development and enculturation on number representation in the brain. *Nature Reviews Neuroscience, 9*, 278–291.
Ansari, D., & Dhital, B. (2006). Age-related changes in the activation of the intraparietal sulcus during nonsymbolic magnitude processing: an event-related functional magnetic resonance imaging study. *Journal of Cognitive Neuroscience, 18*, 1820–1828.
Ansari, D., Garcia, N., Lucas, E., Hamon, K., & Dhital, B. (2005). Neural correlates of symbolic number processing in children and adults. *Neuroreport, 16*, 1769–1773.
Borgmann, K. W. U., Risko, E. F., Stolz, J. A., & Besner, D. (2007). Simon says: Reliability and the role of working memory and attentional control in the Simon Task. *Psychonomic Bulletin & Review, 14*, 313–319.
Cohen Kadosh, R., & Walsh, V. (2007). Dyscalculia. *Current Biology, 17*, 946–947.
Dehaene, S. (1992). Varieties of numerical abilities. *Cognition, 44*, 1–42.
Dehaene, S. (1997). *The number sense*. New York/Cambridge (UK): Oxford University Press/Penguin press.
Dehaene, S. (1996). The organization of brain activations in number comparison: Event-related potentials and the additive-factors method. *Journal of Cognitive Neuroscience, 8*, 47–68.
Dehaene, 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.
Holloway, I. D., & Ansari, D. (2009). Mapping numerical magnitudes onto symbols. *Journal of Experimental Child Psychology, 103*, 17–29.
Holloway, I. D., Price, G. R., & Ansari, D. (2010). Common and segregated neural pathways for the processing of symbolic and nonsymbolic numerical magnitude: An fMRI study. *NeuroImage, 49*, 1006–1017.
Kopriva, R. J., & Shaw, D. G. (1991). Power estimates: The effect of dependent variable reliability on the power of one-factor anovas. *Educational and Psychological Measurement, 51*, 585–595.
Libertus, M., Woldorff, M., & Brannon, E. M. (2007). Electrophysiological evidence for notation independence in numerical processing. *Behavioral and Brain Functions, 3*, 1.
Moyer, R. S., & Landauer, T. K. (1967). Time required for judgments of numerical inequality. *Nature, 215*, 1519–1520.
Murphy, K. R., & Davidshofer, C. O. (2005). *Psychological testing: Principles and applications* (6th ed.). Upper Saddle River, NJ: Pearson Education.
Nieder, A. (2005). Counting on neurons: The neurobiology of numerical competence. *Nature Renews Neuroscience, 6*, 177–190.
Piazza, M., Izard, V., Pinel, P., Le Bihan, D., & Dehaene, S. (2004). Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron, 44*, 547–555.
Santens, S., Roggerman, C., Fias, W., & Verguts, T. (2010). Number processing pathways in human parietal cortex. *Cerebral Cortex, 1*, 77–88.
Schneider, W., Eschman, A., & Zuccolotto, A. (2001). *E-prime user's guide*. Pittsburgh: Psychology Software Tools Inc.
Simon, J. R. (1990). The effects of an irrelevant directional cue on human information processing. In R. W. Proctor & T. G. Reeve (Eds.), *Stimulus–response compatibility: An integrated perspective* (pp. 31–86). Amsterdam: North-Holland.
Stolz, J. A., Besner, D., & Carr, T. H. (2005). Implications of measures of reliability for theories of priming: Activity in semantic memory is inherently noisy and uncoordinated. *Visual Cognition, 12*, 284–336.
Temple, E., & Posner, M. (1998). Brain mechanisms of quantity are similar in 5-year-olds and adults. *Proceedings of the National Academy of Sciences of the United States of America, 95*, 7836–7841.
Van Opstal, F. (2008). Dissecting the symbolic distance effect: Comparison and priming effects in numerical and nonnumerical orders. *Psychonomic Bulletin & Review, 15*, 419–425.
Van Selst, M., & Jolicœur, P. (1994). A solution to the effect of sample size on outlier elimination. *Quarterly Journal of Experimental Psychology, 47a*, 631–650.
Verguts, T., & Fias, W. (2004). Representation of number in animals and humans: A neural model. *Journal of Cognitive Neuroscience, 16*, 1493–1504.
Waechter, Stolz, & Besner (in press). Visual word recognition: On the reliability of repetition priming. *Visual Cognition*.
Zorzi, M., & Butterworth, B. (1999). A computational model of number comparison. In *Paper presented at the 21st annual conference of Cognitive Science Society, Mahwah, NJ*.