Journal of Experimental Psychology: General

ISSN: 0096-3445

2024, Vol. 153, No. 10, 2510–2523
https://doi.org/10.1037/xge0001644

Numerical Comparison Is Spatial—Except When It Is Not

Fraulein Retanal[1], Véronic Delage[1], Evan F. Risko[2], and Erin A. Maloney[1]
[1] School of Psychology, University of Ottawa
[2] Department of Psychology, University of Waterloo

The numerical distance effect (NDE) is an important tool for probing the nature of numerical representation. Across two studies, we assessed the degree to which the NDE relates to one's performance on spatial tasks to investigate the role of spatial processing in numerical comparison and, by extension, numerical cognition. We administered numerical comparison tasks and a variety of tasks thought to tap into different aspects of spatial processing. Importantly, we administered both the simultaneous comparison task and the comparison to a standard task, given claims that the NDEs that arise in these two tasks are different. In both studies, the NDEs elicited when comparing simultaneously presented numbers were more strongly negatively correlated with an individual's performance on the spatial tasks than the NDEs elicited when comparing numbers to a standard. The implications of these data for our understanding of numerical comparison tasks and numerical cognition more generally are discussed.

Public Significance Statement
The numerical distance effect (NDE) is an important metric of numerical representation. Our research reveals that the process in which the NDE is generated varies depending on the numerical comparison task. This finding refines our understanding of numerical representations and highlights the importance of distinguishing between numerical comparison tasks, as they tap into different underlying comparison processes.

Keywords: numerical distance effect, numerical cognition, numerical comparison, simultaneous comparison task, comparison to standard task

Supplemental materials: https://doi.org/10.1037/xge0001644.supp

The numerical distance effect (NDE) is a classic finding in numerical cognition research and has been influential in our understanding of numerical representation (e.g., Dehaene et al., 1993; Maloney et al., 2010). The NDE is obtained using numerical comparison tasks. In these tasks, participants are presented either with one number and are asked to indicate whether that number is numerically smaller than or larger than a standard (e.g., 5) or are presented with two numbers simultaneously and are asked to indicate which represents the larger (or smaller) quantity. When completing these tasks, the NDE emerges. That is, responses are faster and more accurate when comparing two numbers that are separated by a large (relative to a small) numerical distance (e.g., it is easier to compare 7 and 2 than 7 and 6; Dehaene et al., 1990; Moyer & Landauer, 1967). At least in adults, a more "precise numerical representation" is indicated by smaller NDEs (De Smedt et al., 2009, 2013; Holloway & Ansari, 2009), and smaller NDEs correlate with and predict greater mathematics achievement over time (Bugden & Ansari, 2011; Castronovo & Gobel, 2012; De Smedt et al., 2009; Sasanguie et al., 2017). It is believed that basic numerical magnitude processing serves as a necessary scaffold for the acquisition of foundational math skills, such as arithmetic (Holloway & Ansari, 2008). Once the foundational skills have been concretized, the relation between basic numerical

This article was published Online First September 5, 2024.
Aysecan Boduroglu served as action editor.
Fraulein Retanal https://orcid.org/0000-0003-1699-9800
Véronic Delage https://orcid.org/0000-0002-1610-3894
Erin A. Maloney https://orcid.org/0000-0001-5557-0842
Preregistration and the anonymized version of the data presented are available through the Open Science Framework project page at https://osf.io/7dmjy/?view_only=317a221d6ba1475e8f8ce51e9732aed2. The data and ideas presented in the article were previously disseminated at the Canadian Society for Brain, Behaviour, and Cognitive Sciences Meeting, 2022 and Psychonomic Society Annual Meeting, 2022. This article was funded by Natural Sciences and Engineering Research Council of Canada (Grant 211529) awarded to Erin A. Maloney.

Fraulein Retanal played a lead role in data curation, formal analysis, validation, writing–original draft, and writing–review and editing and an equal role in conceptualization, investigation, methodology, and software. Véronic Delage played a lead role in visualization, a supporting role in data curation, formal analysis, writing–original draft, and writing–review and editing, and an equal role in investigation, methodology, and software. Evan F. Risko played a supporting role in formal analysis, supervision, writing–original draft, and writing–review and editing. Erin A. Maloney played a lead role in funding acquisition, investigation, project administration, and resources, a supporting role in writing–original draft and writing–review and editing, and an equal role in conceptualization, formal analysis, and methodology.

Correspondence concerning this article should be addressed to Erin A. Maloney, School of Psychology, University of Ottawa, 136 Jean Jacques Lussier, Vanier Hall, Ottawa, ON K1N 6N5, Canada. Email: erin.maloney@uottawa.ca

magnitude processing and advanced mathematics seems to diminish. Indeed, there is no evidence of a link between the NDE and the complex math skills of professional mathematicians (Hohol et al., 2020).

It is important to note that when discussing the NDE, researchers differentiate between the symbolic NDE, that is, the effect that arises when comparing Arabic digits, and the nonsymbolic NDE, that is, the effect that arises when comparing nonsymbolic representations of quantity, such as sets of squares (e.g., Bugden & Ansari, 2011; Holloway & Ansari, 2009; Maloney et al., 2010). Further, it is important to note that a distinction is often made between the comparison of symbolic representations of single digits (i.e., digits less than 10) and multidigit numbers (digits above 10; see Nuerk et al., 2015, for a comprehensive review). While the NDE can also be observed with multidigit numbers, these comparisons feature influences that are absent for single digits (e.g., unit–decade compatibility). That is, a pair of two-digit Arabic numbers can be defined as unit–decade compatible if both comparisons between tens and units result in the same decision (e.g., 42–57, given that $4 < 5$ and $2 < 7$), but as unit–decade incompatible if the comparisons between tens and units leads to a different decision (e.g., 47–62, given that $4 < 6$ but $7 > 2$). Indeed, participants respond more slowly and less accurately when decades and units are incompatible with the overall decision than when they result in compatible magnitude judgements (see Nuerk et al., 2001, 2005). To complicate matters further, findings in two-digit numbers cannot be generalized to all multidigit numbers (e.g., four-digit comparisons). Meyerhoff et al. (2012) demonstrated that, unlike two-digit comparisons, the comparison of four-digit and six-digit number pairs uses different comparison strategies such as chunking where multidigit numbers are separated into chunks of shorter digit strings. As such, we limit our analyses to symbolic single digit numerical comparisons.

Accounts of the NDE

The NDE has its foundations in research on judgments of perceptual properties of physical objects. It was observed that when making perceptual comparisons, individuals respond quicker and more accurately when the difference between the perceptual properties of objects is greater, as opposed to when the difference is small (Festinger, 1943). Subsequently, Moyer and Landauer extended this finding to the comparison of symbolic numbers, giving rise to the NDE (note that the effect was only later dubbed the "numerical distance effect" by Dehaene et al., 1990). In their seminal work, Moyer and Landauer (1967), proposed the psychophysical model, in which the effect, known today as the NDE, resulted from an analog representation of numerical magnitude. According to this account the NDE is thought to reflect the subjective distances between the analog representations of digits' magnitude (Moyer & Landauer, 1967). When there is a smaller magnitude difference between two digits, this corresponds to a smaller difference between their analog representations. Consequently, this reduces discriminability between the analog representation leading to greater difficulty (longer response times [RTs] and higher error rates) in comparing the two digits. Since then, this model of an analog representation of magnitude has been extended to take the form of a spatially oriented mental number line where numbers are represented in their numerical order along a continuum (Dehaene et al., 1990, 1993; Restle, 1970; Sekuler & Mierkiewicz, 1977). It is now believed that the orientation of the number line follows that of the participants' dominant language. Specifically, in cultures where one reads left to right, the numbers are represented with smaller quantities on the left and larger quantities on the right (Göbel et al., 2011). Currently, there are two dominant theoretical accounts of the NDE that use this spatially oriented mental number line. According to these accounts, the *representational overlap account* and *working memory account*, the to-be-compared digits are located on a spatial mental number line and as the numeric distance decreases the difficulty of discriminating their location increases, generating the NDE (Restle, 1970; van Dijck & Fias, 2011). The critical difference between these accounts is that, according to the *representational overlap account* (Restle, 1970), the number line is a permanent representation in long-term memory (LTM), whereas in the *working memory account* (van Dijck & Fias, 2011), the number line is stored in working memory (WM) and is only generated in service of completing the task.

According to these accounts, numbers that are closer together on the mental number line share more representational overlap than numbers that are farther apart, leading to slower and less accurate comparisons for numbers that are closer together (i.e., the NDE). As such, a more spatially precise representation (i.e., equal interval spacing between the numbers in the mental number line) should result in a smaller NDE, that is, faster and more accurate responses across all levels of numerical distance. Thus, individuals with smaller NDEs are thought to have a more spatially precise representation of the mental number line than individuals with larger NDEs (Bugden & Ansari, 2011; De Smedt et al., 2009, 2013; Holloway & Ansari, 2009).

A third account of the NDE, the *response competition account* (Verguts et al., 2005), unlike the previous two accounts of the NDE, does not implicate a spatially oriented analog representation. Rather, according to the *response competition account* (Verguts et al., 2005), individuals have a mental store of basic "number facts," akin to a mental lexicon, wherein each digit is differentially associated with the responses "smaller than" and "larger than." As digit magnitude increases, the association between that digit and the "larger than" response increases and the association with the "smaller than" response decreases (see the *discrete semantic system* for a similar approach, Krajcsi et al., 2016). According to the *response competition account*, the NDE results from a greater amount of response competition between two digits with similar strengths of association with the relevant response ("larger than" or "smaller than").

The *response competition account* put forth by Verguts et al. (2005) does not require that numbers be represented spatially. It is worth noting, however, that a spatially defined response association has been put forth to account for the spatial numerical association of response codes effect (Gevers et al., 2006). Accordingly, the left hand is associated with the "smaller than" response and the right hand is associated with the "larger than" response. While the spatial numerical association of response codes and NDE can both be explained by response competition related to associations with "smaller than" and "larger than," the *response competition account* posited by Verguts et al. (2005) does not, at present, necessitate a role for spatial information in the generation of the NDE.

In summary, the NDE in symbolic comparison is an important metric of the representation and processing of numerical magnitude. However, the underlying nature of the NDE, and whether or not is it rooted in a spatial representation, as implied within the *representational overlap account* (Restle, 1970) and *working memory account*

(van Dijck & Fias, 2011) but not the *response competition account* (Verguts et al., 2005), remains an empirical question.

Divergence of the Two Numerical Comparison Tasks

To complicate matters, recent research has shown that the two numerical comparison tasks—which have been used interchangeably to date—may not be tapping into the same numerical comparison strategy (Maloney et al., 2010, 2019). Specifically, the NDE generated in the simultaneous comparison task decreases under a verbal WM load while the NDE in the comparison to a standard task increases under the same load (Maloney et al., 2019). To explain this result, Maloney et al. (2019) suggested that participants could adopt either a long-term or short-term comparison strategy in successive comparison tasks. Maloney et al. (2019) argue that, in the simultaneous comparison task, the digits are compared in WM and that in the comparison to a standard task, on each trial, participants retrieve the comparison digit (5) from long-term memory to carry out the comparison. This suggests that while both numerical comparison tasks generate an NDE, they may not be tapping into the same representation or processes. The idea that the two tasks are not tapping into the same representation or processes is consistent with the finding that the NDEs that arise within the two tasks are not strongly correlated (Maloney et al., 2010).

The Present Study

Building upon the *representational overlap account* (Restle, 1970) and the *working memory account* (van Dijck & Fias, 2011), which posit that the comparison of small numbers is occurring using a mental number line that is spatially represented, we propose here a novel account of the relation between the NDE and spatial competencies. Specifically, we theorize that individuals who have higher spatial competency can summon a spatial number line that is more precise (i.e., equal interval spacing between the numbers in the mental number line) than their peers with lower spatial competency. Individuals with a more spatially precise mental number line should have more precise representations of number (i.e., less representational overlap) and therefore exhibit smaller NDEs. As such, individuals with higher spatial competencies should exhibit smaller NDEs. The novel addition here is the proposed role of spatial competency in the generation of the mental number line and the hypothesized link between spatial competency and the NDE. While both the *representational overlap account* (Restle, 1970) and the *working memory account* (van Dijck & Fias, 2011), posit a spatially represented mental number line, neither account outlines the way in which individual differences in spatial competency link to numerical representation and the NDE.

In contrast, the *response competition account* (Verguts et al., 2005) posits an alternative perspective. According to the *response competition account* (Verguts et al., 2005), the NDE arises because of a greater amount of response competition between two digits with similar strength of associations with the relevant response (i.e., larger than vs. smaller than). Because the mechanism that gives rise to the NDE, according to the *response competition account*, is not posited to be spatial in nature, there is no reason to predict, based on this account, that one's spatial competency would be related to the size of their NDE. Here, we have derived predictions that focus on the coarse presence versus absence of a correlation between the NDE and spatial competencies. At present, this is the strongest prediction one could derive based on the available literatures. Future theoretical development will hopefully lead to theories that make stronger predictions regarding the strength of the correlations.

To test the competing predictions that can be derived from the *representational overlap* and the *working memory accounts* of the NDE, and *response competition account* of the NDE, we administered numerical comparison tasks and a variety of tasks thought to tap into different aspects of spatial processing. It is unclear which spatial skills are implicated within the accounts of the NDE. Indeed, the term "spatial" is often used as an umbrella term, referring to multiple related but distinct skills (Uttal et al., 2013). These various skills can be indexed using different types of tasks (e.g., Uttal et al., 2013). However, there remains debate within the field regarding how to divide, define and name these subcategorizes of spatial skills (e.g., Carroll, 1993; Höffler, 2010; Linn & Petersen, 1985; Mix & Cheng, 2012; Mix et al., 2018; Uttal et al., 2013). While a widely accepted model for categorizing spatial skills uses a multidimensional model proposed by Uttal et al. (2013), recent research suggests a developmental trend in which the dimensions of how spatial skills are categorized converge (i.e., less dimensionality; Mix et al., 2018). Further, there is a dearth of research investigating the relation between spatial skills and the NDE. Previous studies have shown that the NDE in a comparison to standard task is not related to 2D mental rotation performance (Viarouge et al., 2014) and does not rely on spatial working memory resources (Herrera et al., 2008). However, these studies did not explore how the NDE elicited by the simultaneous comparison task may relate to spatial processing.

Given the lack of consensus on the definition and categorization of spatial skills, as well as the absence of a strong foundation linking specific spatial skills to the NDE we decided to employ a variety of spatial tasks and create a composite variable, "spatial index," representing spatial competency. In this context, we use the term "spatial competency" to encompass a general representation of spatial skills, without specifying individual spatial skills. While we predict a relation between the NDE and spatial competencies, we remain agnostic as to whether the NDE will relate to performance on all spatial tasks or only certain spatial tasks. Further, we opted to administer both variants of the numerical comparison task—the simultaneous comparison task and the comparison to a standard task—given claims that the NDEs that emerge from these two tasks may not share the same underlying mechanism (Maloney et al., 2010, 2019).

Study 1

Methodology

Participants

Two hundred forty-nine participants were recruited from an undergraduate student research pool and completed all measures. After the data cleaning (described in the scoring section below), the final sample size was 227 participants. Participants were asked what their gender is and had the option of selecting: "Male," "Female" or "You don't have an option that applies to me. I identify as (please specify)." Those who selected the last option were invited to write the term that best described their gender identity. Of the included participants, 165 identified as female, 59 identified as male, one identified as genderqueer, and two identified as nonbinary.

Participants reported their age in a free-response box. The mean age of the sample was 19.73 years of age (*SD* = 3.33, 17–40 years old). One participant did not report their age.

Procedure

In Study 1, participants completed measures of spatial performance, numerical comparison, and working memory. To assess participant's spatial skills, they completed four separate tasks: dot localization, mental rotation, navigation, and perspective taking. Participants also completed a simultaneous comparison task, a comparison to a standard task, and a backward letter span task (to assess WM capacity). The order of the spatial tasks, numerical comparison tasks, and backward letter span task were counterbalanced, and a demographics questionnaire was completed last. All tasks and questionnaires were completed online using the Gorilla platform, in English.

Measures

Backward Letter Span. A host of factors may play a role in the relation between the numerical comparison tasks and the spatial tasks. We attempt to address this in one way using a task that indexes WM capacity as a starting point. Given the characterization of WM as a "dissociable cognitive skill with unique links to learning outcomes" (Alloway & Alloway, 2010, p. 26), we included it as a domain general measure to help control for the possibility that people who are strong in one academic area (e.g., numerical comparison) are simply likely to be strong in all academic areas (e.g., spatial tasks). Participant's working memory capacity was measured using a computerized version of the backward letter span task (Maloney et al., 2009; Wechsler, 1997). Participants were shown a series of letters at a rate of 1 letter per second. At the end of the sequence, participants had to recall the letters shown in backwards order. The series of letters ranged from two letters to eight, starting with two letters and increasing by one letter every two trials for a total of 14 trials. Scoring was discontinued when the participants failed both trials of a given number of letters. Scoring was based on the highest number of letters in a series correctly recalled by the participant. Thus, scoring ranged from 0 to 8.

Numerical Comparison Tasks. Participants completed two numerical comparison tasks—the simultaneous comparison task and the comparison to a standard task. Each task began with five practice trials where instructions remained on screen until dismissed by the participant. Each task consisted of 80 trials, with rests in between where participants were asked to press <SPACEBAR> to proceed to the next task.

Comparison to a Standard. Participants were instructed to compare the to-be presented digits to 5 (the standard). Each trial began with a fixation sign that remained on the screen for 3s. Afterward, a display containing an Arabic digit was presented. Numbers were 1, 4, 6, or 9. The numerical distance between the stimuli was 1 or 4, with an equal number of trials at each distance. Participants were told to press the "A" key if the presented number was lower, and the "L" key if it was higher than 5. Participants pressed <SPACEBAR> to proceed to the next trial.

Simultaneous Comparison. Participants were instructed to indicate which of the two to-be presented digits were numerically larger. Each trial began with a fixation sign that remained on the screen for 3s. Afterward, a display containing two Arabic digits was presented. Numbers ranged from 1 to 4 and from 6 to 9. The numerical distance between the stimuli was 1 or 4, with an equal number of trials at each distance. Participants were told to identify which of the two numbers was numerically larger by pressing the "A" key if the left number was larger, and the "L" key if the right number was larger. Participants pressed <SPACEBAR> to proceed to the next trial.

Spatial Tasks. The four spatial tasks selected have demonstrated good reliability in the previous research (Cronbach's α range: .89–.92; Daker et al., 2022). Spatial tasks were analyzed individually and as a composite measure, the spatial index. To create the spatial index, we first created a standardized *z*-score for accuracy of each spatial task and then created an average of each of the four standardized scores.

Dot Localization. Participants' dot localization performance was measured using 15 trials of a modified computerized version of the dot localization task (Manna et al., 2010). In this task, participants were first presented with a rectangle containing two dots for 125 ms. Once this rectangle had disappeared, the participant was presented with another rectangle containing a grid and was asked to press locations within the grid to identify where the two dots, previously shown, would have been located if both rectangles had been superimposed. Participants obtained a score of 1 on each item if they correctly identified the location of both dots, a score of 0 on each item if they incorrectly identified one or both dots, for a maximum score of 15.

Mental Rotation. Participants' mental rotation performance was measured using 15 trials of a computerized mental rotation task (Shepard & Metzler, 1971). Participants were shown two 3D objects made of 10 adjoining cubes which were oriented in different directions. In this task, participants were asked to identify whether they thought the two objects shown were the same objects oriented differently or if they were two different objects by clicking "Same" or "Different" on the screen. Participants obtained a score of 1 for correct items and 0 for incorrect items for a maximum score of 15.

Navigation. Participants' navigation performance was measured using a modified computerized version of the Road-Map Test of Directional Sense (Ferguson et al., 2015; Money et al., 1965). In this task, participants were presented a map that contained a dotted path. On each "street" corner, participants were shown the letter R (right turn) or L (left turn) to demonstrate the direction they would be turning if they were walking along the dotted path. However, not every turn was labeled correctly. Therefore, participants were required to press "Y" or "N" to identify whether they agreed or disagreed with the direction provided. Three maps were presented with three, 17 and 33 turns. Consistent with the previous studies (e.g., Ferguson et al., 2015), only the third map with 33 turns was used to score performance, the other two were used as practice trials. Participants obtained a score of 1 for each correctly identified turn of the 33-turn map for a maximum of 33 points.

Perspective Taking. Participants' perspective-taking performance was measured using 15 trials of the Hegarty perspective-taking task (Hegarty & Waller, 2004). In this task, participants were shown a screen with a variety of common objects (cat, car, house, etc.) and an "arrow circle." The participants were asked to imagine that they were standing in the location of the object in question (object A in the middle of the circle) and facing a particular point (object B at the top of the circle). They were then asked to determine

in which direction they would find a third object (object C) by selecting one of the five areas around the circle, each area was identified with a letter (a–e). Participants obtained a score of 1 for correct items and 0 for incorrect items for a maximum score of 15.

Scoring

Accuracy and RTs were collected for the numerical comparison tasks, the spatial tasks and the backward letter span. The response times of the numerical comparison tasks were used to calculate the NDEs for each participant and the accuracy (% correct) of the spatial tasks was used to calculate the spatial index score. On the spatial tasks, when a nonresponse was recorded for an item (i.e., 1 data point was missing), the participant was given a score of 0 on the trial (i.e., the trial counted as an incorrect trial). We removed two participants from the analysis for using the same response category, consecutively, on one or more of the measures, identified by using longstring analysis (Yan, 2008) on the spatial task. RTs and accuracies on the numerical comparison tasks were then analyzed across the 247 participants remaining participants.

Comparison to Standard Task. For the comparison to standard task, we excluded participants below chance accuracy for the numerical comparison task (i.e., accuracy below 50%), resulting in the removal of four participants trials on which participants responded incorrectly to the numerical comparison task (3.4%) and trials that were too fast (<200 ms) and too slow (>5,000 ms; 1.0%) were removed. Further, trials were removed for having RTs greater than or equal to 2.5 *SD*s away from the participant's mean RT at each numerical distance (3.1% of remaining trials). The NDE was calculated by subtracting the participant's mean RTs at distances of 4 from mean RTs at distances of 1. Last, four participants with NDEs greater than or equal to 2.5 *SD*s away from the sample mean were removed. This left us with 239 participants for this task.

Simultaneous Comparison Task. For the simultaneous comparison task, we excluded participants chance accuracy for the numerical comparison tasks (i.e., accuracy below 50%), resulting in removal of seven participants. Trials on which participants responded incorrectly to the numerical comparison task (3.0%) and trials that were too fast (<200 ms) and too slow (>5,000 ms; 1.1%) were removed. Further, trials were removed for having RTs greater than or equal 2.5 *SD*s away from the participant's mean RT at each numerical distance (2.8% of remaining trials). The NDE was calculated by subtracting the participant's mean RTs at distances of 4 from mean RTs at distances of 1. Last, nine participants with NDEs greater than or equal to 2.5 *SD*s away from the sample's mean NDE were removed. This left us with 231 participants for this task.

Although all participants completed both numerical comparison tasks, the exclusion criteria that were applied separately left us with an unequal number of participants for each numerical comparison task. To ensure that the sample size is consistent when comparing correlations, only the 227 participants who were not excluded from either of the two numerical comparison tasks were included in the analyses. A post hoc analysis was performed using G*Power Version 3.1.9.7 (Faul et al., 2007) to determine the smallest detectable effect size for bivariate correlations, with power set at 0.8 and our sample size ($N = 227$). The analysis indicated that a correlation as low as $r = 0.185$ can be detected with our current sample size and desired power level.

Transparency and Openness

We report all data exclusions, all manipulations, and all measures in the study. All the data and the syntax needed to reproduce analyses for Study 1 are available at https://osf.io/7dmjy/ Data were analyzed using SPSS software and R and the package Estimate Split-Half Reliabilities (Pronk et al., 2022).

Results

See Figure 1 for a visual representation of the density plots, scatterplots, and Pearson's correlations for the NDE of the numerical comparison tasks and accuracy of the spatial tasks.

Reliabilities

The reliability of the NDE (RT) generated in each numerical comparison task was calculated using a stratified permuted split method using the R package Estimate Split-Half Reliabilities (Pronk et al., 2022). The data were randomly split to calculate two NDE (RT) scores for each participant, this procedure was replicated 1,000 times. The stratification ensured that each split had an equal number of trials with a numerical distance of 1 and 4. Spearman–Brown corrected Pearson correlations were calculated for each split. The average of the coefficients was calculated to obtain the reliability estimates. The reliability estimates of the spatial tasks' accuracy were calculated using the same method, without stratification. The reliability estimates and their nonparametric bias-corrected and accelerated bootstrap 95% confidence intervals are presented in Table 1. We also include the histograms of the Spearman–Brown corrected Pearson correlations of the 1,000 replications for both numerical comparison tasks in Figure S1 of the Supplemental Materials.

We observed acceptable reliabilities for all four spatial tasks (r_{SB} = 0.62–0.98) and poor reliability for the NDE in the simultaneous comparison task ($r_{SB} = 0.15$) and the comparison to standard task ($r_{SB} = 0.30$). This is critical because the poor reliability of the numerical comparison tasks attenuate the observed correlations. Thus, it is important to keep these reliabilities in mind in interpreting the results presented below.

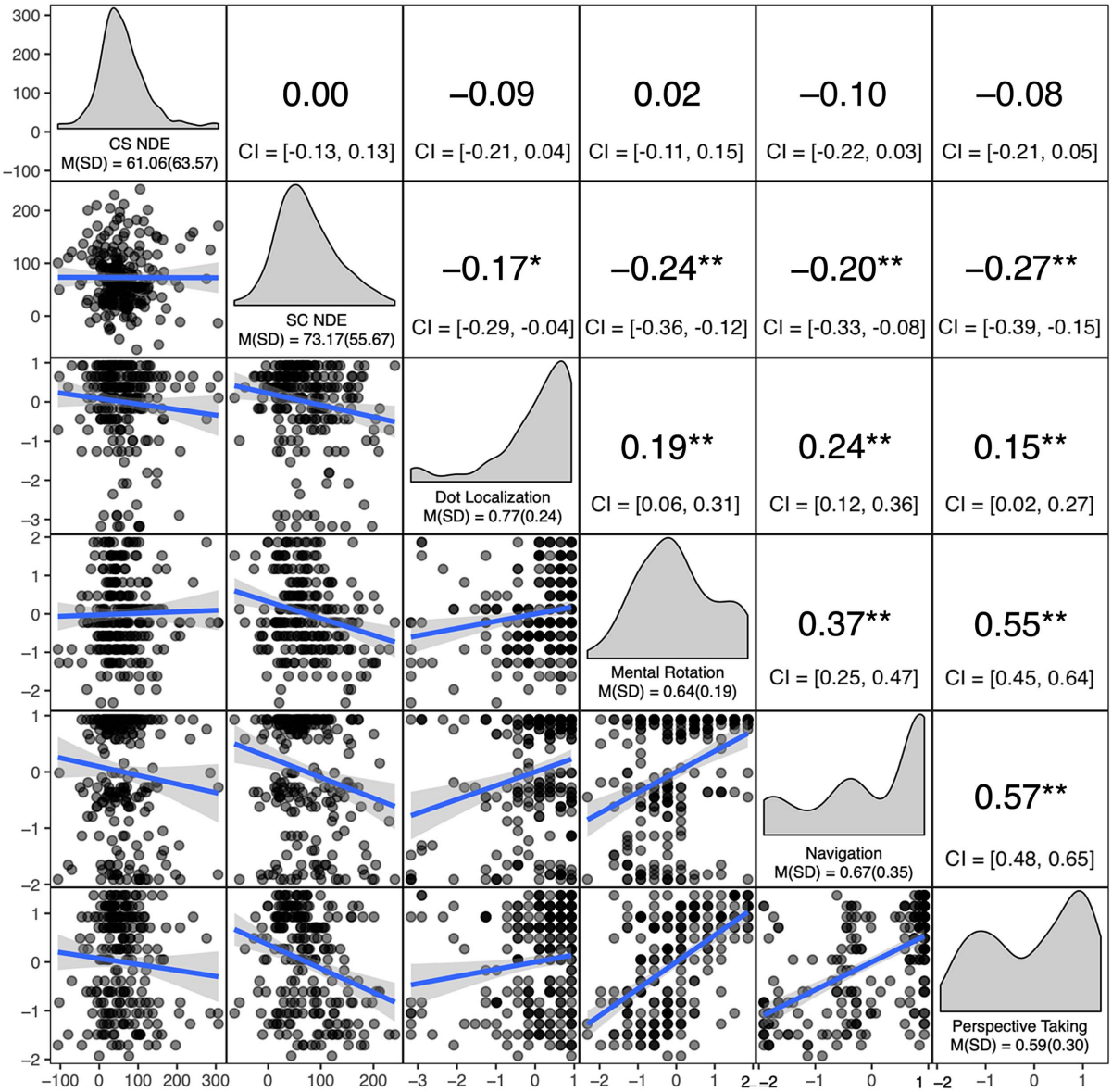
Numerical Distance Effects

We first examined whether there was a significant NDE in each of the two comparison tasks. As was expected, there was a significant NDE in RTs for correct trials in both the simultaneous comparison task, $t(226) = 19.80$, $p < .001$, $d = 1.31$, such that responses were faster at distances of 4 (647 ms) than for distances of 1 (721 ms), and in the comparison to a standard task, $t(226) = 14.47$, $p < .001$, $d = .96$, such that responses were faster at distances of 4 (636 ms) relative to distances of 1 (697 ms).

Parallel analyses were conducted on the accuracy data. Again, there was a significant NDE in the simultaneous comparison task, $t(226) = 9.09$, $p < .001$, $d = 0.55$, such that responses were more accurate when comparing at distances of 4 (1.5% error) relative to distances of 1 (4.0% error) and the comparison to a standard task, $t(226) = 8.35$, $p < .001$, $d = 0.60$, such that responses were more accurate at distances of 4 (2.2% error) relative to distances of 1 (4.2% error).

Figure 1

Density Plots, Scatterplots, and Pearson's Correlations for Study 1



Note. Density plots of numerical comparison and spatial tasks are depicted on the diagonal, scatterplots with linear trends and 95% confidence intervals for each bivariate association are depicted below the diagonal, and Pearson's correlation coefficients and confidence intervals for each bivariate association are depicted above the diagonal. CS NDE = comparison to a standard numerical distance effect, reaction time, SC NDE = simultaneous comparison numerical distance effect, reaction time. Accuracies (*z*-scored) of spatial tasks are shown in the scatterplot. CI = confidence interval. See the online article for the color version of this figure.

* $p < .05$. ** $p < .01$.

Importantly, although the NDEs were similar in magnitude (simultaneous comparison task = 74 ms; comparison to a standard task = 61 ms; simultaneous comparison task error rate = 0.02; comparison to a standard task error rate = 0.02), the NDEs generated in the tasks were not significantly correlated in RTs, $r(225) = 0.00$ $p = .97$ CI [−0.13, 0.13], or errors $r(225) = -0.05$, $p = .49$. These findings parallel those of previous findings suggesting that the NDEs elicited in the two tasks are not the same (Maloney et al., 2010).

Spatial Tasks

The four spatial tasks were significantly intercorrelated ($r = 0.15$ to 0.57). Consistent with the *representational overlap account*

Table 1
Reliabilities of the NDE From Numerical Comparison Tasks and the Spatial Tasks for Study 1

Task	Reliability (*SD*)	95% confidence interval
Comparison to standard NDE (RT)	0.30 (0.15)	[0.04, 0.48]
Simultaneous comparison NDE (RT)	0.15 (0.21)	[−0.26, 0.37]
Dot localization (accuracy)	0.86 (0.01)	[0.81, 0.90]
Mental rotation (accuracy)	0.62 (0.04)	[0.54, 0.68]
Navigation (accuracy)	0.98 (0.003)	[0.97, 0.98]
Perspective taking (accuracy)	0.88 (0.01)	[0.86, 0.90]

Note. Reliability estimates are the average of Spearman–Brown corrected Pearson correlation coefficients of 1,000 replications. Confidence intervals are nonparametric bias-corrected and accelerated bootstrap 95% confidence intervals for the coefficients averaged across split replications. NDE = numerical distance effect; RT = response times.

(Restle, 1970) and the *working memory account* (van Dijck & Fias, 2011), all four spatial tasks were significantly correlated with the NDE in the simultaneous comparison task (see Figure 1). That is, as performance on the spatial tasks increased, the size of the NDE in simultaneous comparison task decreased (indicative of a more precise numerical representation). Furthermore, the pattern remained even after controlling for performance on the backwards letter span task: dot localization, $r(224) = -0.19$, $p < .05$, CI [−0.31, −0.06], mental rotation, $r(224) = -0.23$, $p < .001$, CI [−0.35, −0.10], navigation, $r(224) = -0.16$, $p < .05$, CI [−0.28, −0.03], and perspective taking, $r(224) = -0.24$, $p < .001$, CI [−0.36, −0.11]. In contrast, and consistent with the *response competition account* (Verguts et al., 2005), the correlation between the four spatial tasks and the NDE for comparison to standard were not significant (see Figure 1). Importantly, the pattern remained after controlling for performance on the backwards letter span task: dot localization, $r(224) = -0.09$, $p = .20$, CI [−0.21, 0.04]; mental rotation, $r(224) = 0.02$, $p = .79$, CI [−0.11, 0.15]; navigation, $r(224) = -0.12$, $p = .07$, CI [−0.25, 0.01]; and perspective taking, $r(224) = -0.10$, $p = .15$, CI [−0.22, 0.03].

Spatial Index

We next examined whether the NDE generated during a numerical comparison task was significantly related to the spatial index. There was a significant correlation between the NDE generated in the simultaneous comparison task and the spatial index, $r(225) = -0.31$, $p < .01$, CI [−0.42, −0.19], such that those who performed better on the spatial index had smaller NDEs in the simultaneous comparison task. This relation remained even after controlling for performance on the backwards letter span task, $r(224) = -0.28$, $p < .001$, CI [−0.40, −0.16]. There was no significant relation between the NDE generated in the comparison to a standard task and performance on the spatial index, $r(225) = -0.08$, $p = .21$ CI [−0.21, 0.05]. This remained the case even after controlling for performance on the backwards letter span task, $r(224) = -0.10$, $p = .14$, CI [−0.23, 0.03].

To directly compare correlations between the spatial index and the NDE of the two numerical comparison tasks, we used Zou's (2007) approach for dependent overlapping correlations. The correlations were compared using cocor, a web interface for the statistical comparison of correlations (Diedenhofen & Musch, 2015). The correlation between the spatial index and the NDE of the simultaneous comparison task was significantly larger than the correlation between the spatial index and the NDE of the comparison to standard task, as indicated by the fact that the 95% CI of the difference between the two correlations did not cross zero (Zou, 2007), 95% CI [0.05, 0.40]. Note that we opted to only compare the size of the correlations between the NDEs and the spatial index, and not each spatial task individually to limit the overall number of analyses conducted.

The NDEs observed in the accuracy data did not correlate with the spatial index in either task, simultaneous comparison $r(225) = -0.09$, $p = .19$; comparison to standard $r(225) = 0.03$, $p = .67$.

Discussion

In Study 1, we assessed the degree to which the two NDEs related to an individual's spatial skills to investigate the role of spatial processing in numerical comparison. The NDE elicited when comparing simultaneously presented numbers was more strongly related to performance on the spatial tasks than was the NDE elicited when comparing numbers to a standard. That is, smaller NDEs were related to better spatial performance for the simultaneous comparison task and significantly less so for the comparison to a standard task. The presence of a significant correlation between the NDE and performance on the spatial index is in line with the novel theory, tested herein, that the ability to summon a more spatially precise representation of a spatially oriented mental number line (as measured by spatial competence) would result in a smaller NDE. Notably, this pattern was observed more strongly for the NDE elicited by the simultaneous comparison task. This dissociation between the two comparisons tasks is consistent with the claim that there are different mechanisms within the two numerical comparison tasks that give rise to the NDE (Maloney et al., 2010, 2019). Indeed, these data appear to provide compelling evidence of a stronger relation between the NDE generated within the simultaneous comparison task and performance on the spatial index than that generated within the comparison to a standard task. However, it is critical to note that the split-half reliabilities of the NDES generated within both tasks are poor.

Given poor split-half reliability of a measure attenuates the observed correlation between that measure and other measures (Spearman, 1910), and the novelty of these findings, we felt it prudent to conduct a replication study. Before conducting Study 2, we preregistered our methods and analyses; see (https://osf.io/fzx3y; Retanal et al., 2022).

Study 2

Methodology

Participants

Two hundred fifty participants were recruited from Prolific, an online participant recruitment platform but only 248 completed all measures. After the data trimming procedure (described in the scoring section below), the final sample size was 223 participants. Participants were asked what their gender is and had the option of selecting: "Male," "Female" or "You don't have an option that applies to me.

I identify as (please specify)." Those who selected the last option were invited to write the term that best described their gender identity. Of the included participants, 113 identified as female, 103 as male, four as nonbinary, one as gender fluid, one as demiwomen, and one participant did not respond to the question. Participants reported their age in a free-response box. The mean age of the participants was 25.86 years of age ($SD = 6.88$, 18–56 years old). One participant did not report their age.

Procedure

In Study 2, participants completed the same measures of spatial performance, numerical comparison, and working memory as in Study 1. The order of the tasks was counterbalanced, and a demographics questionnaire was completed last. All tasks and questionnaires were completed online using the Gorilla platform, in English.

Measures

The same measures were used as in Study 1.

Scoring

In Study 2, we employed the identical data scoring and trimming procedure to that used in Study 1. We removed two participants for using the same response category, consecutively, on one or more of the spatial measures. Next, RTs and errors on the numerical comparison tasks were analyzed across the 246 remaining participants.

Comparison to Standard Task. For the comparison to a standard task, we excluded participants below chance accuracy for the numerical comparison tasks (i.e., accuracy below 50%), resulting in the removal of seven participants. Trials on which participants responded incorrectly to the numerical comparison task (2.3%) and trials that were too fast (<200 ms) and too slow (>5,000 ms; 0.1%) were removed. Further, trials in were removed for having RTs greater than or equal to 2.5 *SD*s away from the participant's mean RT at each numerical distance (2.9% of remaining trials). The NDE was calculated by subtracting the participant's mean RTs at distances of 4 from mean RTs at distances of 1. Last, nine participants with NDEs greater than or equal to 2.5 *SD*s away from the sample mean were removed. This left us with 230 participants for this task.

Simultaneous Comparison Task. For the simultaneous comparison task, we excluded participants below chance accuracy for the numerical comparison tasks (i.e., accuracy below 50%), resulting in the removal of one participant. Trials on which participants responded incorrectly to the numerical comparison task (2.5%) and trials that were too fast (<200 ms) and too slow (>5,000 ms; 0.2%) were removed. Further, trials were removed for having RTs greater than or equal 2.5 *SD*s away from the participant's mean RT at each numerical distance (2.5% of remaining trials). The NDE was calculated by subtracting the participant's mean RTs at distances of 4 from mean RTs at distances of 1. Last, seven participants with NDEs greater than or equal to 2.5 *SD*s away from the sample's mean NDE were removed. This left us with 238 participants for this task.

To ensure that the sample size is consistent when comparing correlations, the 223 who were not excluded from either of the two numerical comparison tasks were included in the analyses. A post hoc analysis was performed using G*Power Version 3.1.9.7 (Faul et al., 2007) to determine the smallest detectable effect size for bivariate correlations, with power set at 0.8 and our sample size ($N = 223$). The analysis indicated that a correlation as low as $r = 0.1863$ can be detected with our current sample size and desired power level.

Transparency and Openness

We report all data exclusions, all manipulations, and all measures in the study. The research design and analysis plan for Study 2 was preregistered; see https://doi.org/10.17605/OSF.IO/FZX3Y. All the data and the syntax needed to reproduce analyses for Study 2 are available at https://osf.io/7dmjy/. Data were analyzed using SPSS and R and the R package Estimate Split-Half Reliabilities (Pronk et al., 2022).

Results

See Figure 2 for a visual representation of the density plots, scatterplots, and Pearson's correlations for the RT NDE of the numerical comparison tasks and accuracy of the spatial tasks.

Reliabilities

The reliabilities of the NDEs generated within the numerical comparison tasks (RT) and the spatial tasks (accuracy) were calculated in the same way as in Study 1. The reliability estimates and their confidence intervals are presented in Table 2. Histograms of the Spearman–Brown corrected Pearson correlations of the 1,000 split replications for both numerical comparison tasks are in Figure S2 of the Supplemental Materials.

We observed acceptable reliabilities for all four spatial tasks ($r_{SB} = 0.58$–0.96). Although still poor, we observed higher reliabilities than in Study 1 for the NDE in the simultaneous comparison task ($r_{SB} = 0.44$) but not for the NDE in the comparison to standard task ($r_{SB} = 0.25$).

Numerical Distance Effect

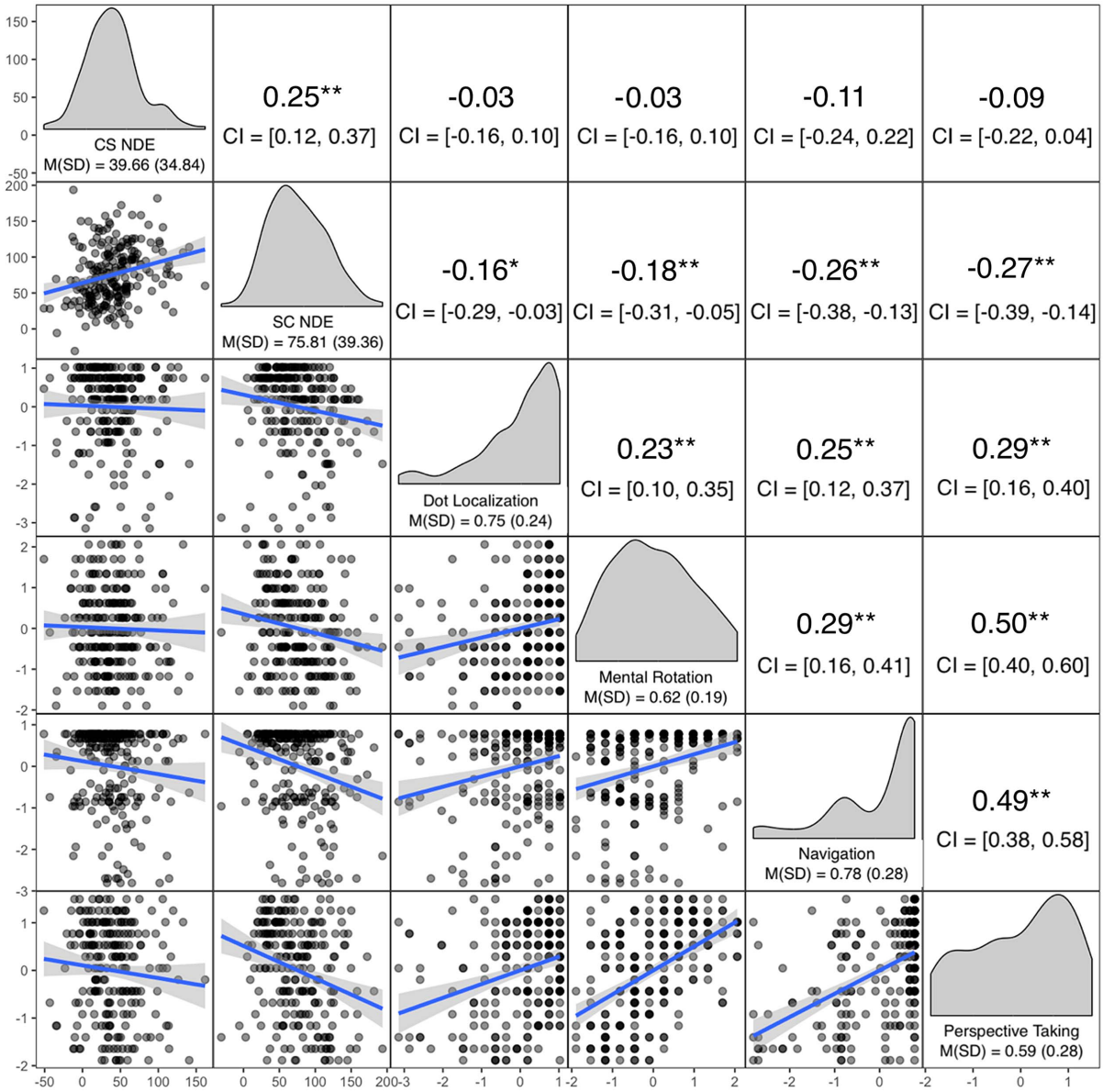
Similar to Study 1, there was a significant NDE in RTs for correct trials in both the simultaneous comparison task, $t(222) = 28.65$, $p < .001$, $d = 1.91$, such that responses were faster for distances of 4 (557 ms) than for distances of 1 (632 ms), and in the comparison to a standard task, $t(222) = 17.00$, $p < .001$, $d = 1.14$, such that responses were faster for distances of 4 (519 ms) than for distances of 1 (559 ms).

Parallel analyses were conducted on the accuracy data. Again, there was a significant NDE in the simultaneous comparison task, $t(222) = 14.36$, $p < .001$, $d = 0.96$, such that responses were more accurate when comparing at distances of 4 (0.5% error) relative to distances of 1 (4.4% error), and the comparison to a standard task, $t(222) = 9.76$, $p < .001$, $d = 0.65$, such that responses were more accurate when comparing at distances of 4 (1.1% error) relative to distances of 1 (3.6% error).

Unlike the relation between the NDE in the simultaneous comparison task and the comparison to standard task found in Study 1, the NDE generated in the two numerical comparison tasks

Figure 2

Density Plots, Scatterplots, and Pearson's Correlations for Study 2



Note. Density plots of numerical comparison and spatial tasks are depicted on the diagonal, scatterplots with linear trends and 95% confidence intervals for each bivariate association are depicted below the diagonal, and Pearson's correlation coefficients and their 95% confidence intervals for each bivariate association are depicted above the diagonal. CS NDE = comparison to a standard numerical distance effect, reaction time, SC NDE = simultaneous comparison numerical distance effect, reaction time. Accuracies (*z*-scored) of spatial tasks are shown in the scatterplots. CI = confidence interval. See the online article for the color version of this figure.

* $p < .05$. ** $p < .01$.

were significantly correlated in RTs, $r(222) = 0.25$, $p < .01$, CI [0.12, 0.37] and errors, $r(225) = 0.20$, $p < .01$.

Spatial Tasks

Like Study 1, the four spatial tasks were significantly intercorrelated ($r = 0.23$–0.50). In addition, all four spatial tasks were significantly correlated with the NDE in the simultaneous comparison task (see Figure 2). That is, as performance on the spatial tasks increased, the size of the NDE in simultaneous comparison task decreased (indicative of a more precise numerical representation). Furthermore, the pattern remained even after controlling for performance on the backwards letter span task: dot localization, $r(220) = –0.14$, $p < .05$, CI [−0.26, −0.01]; mental rotation, $r(220) = –0.15$, $p < .05$, CI [−0.28, −0.02]; navigation, $r(220) = –0.23$, $p < .001$, CI [−0.35, −0.10]; and perspective

Table 2
Reliabilities of the NDE From Each Numerical Comparison Task and the Spatial Tasks for Study 2

Task	Reliability (*SD*)	95% confidence interval
Comparison to standard NDE (RT)	0.25 (0.14)	[0.01, 0.41]
Simultaneous comparison NDE (RT)	0.44 (0.16)	[0.02, 0.57]
Dot localization (accuracy)	0.84 (0.02)	[0.79, 0.88]
Mental rotation (accuracy)	0.58 (0.04)	[0.50, 0.64]
Navigation (accuracy)	0.96 (0.00)	[0.95, 0.97]
Perspective taking (accuracy)	0.84 (0.01)	[0.81, 0.87]

Note. Reliability estimates are the average of Spearman–Brown corrected Pearson correlation coefficients of 1,000 split replications. Confidence intervals are nonparametric bias-corrected and accelerated bootstrap 95% confidence intervals for the coefficients averaged across split replications. NDE = numerical distance effect; RT = response times.

taking, $r(220) = -0.23$, $p < .001$, CI [−0.35, −0.10]. Similar to Study 1, the correlation between the four spatial tasks and the NDE for comparison to standard were not significant (see Figure 2). Importantly, this did not change after controlling for performance on the backwards letter span task: dot localization, $r(220) = -0.02$, $p = .73$, CI [−0.15, 0.10]; mental rotation, $r(220) = 0.02$, $p = .73$, CI [−0.15, 0.11]; navigation, $r(220) = -0.10$, $p = .12$, CI [−0.23, 0.03]; and perspective taking, $r(220) = -0.09$, $p = .19$, CI [−0.22, −0.04].

Spatial Index

Replicating what was observed in Study 1, there was a significant correlation between the NDE in simultaneous comparison and the spatial index, $r(221) = -0.31$, $p < .01$, CI [−0.42, −0.18], which remained even after controlling for performance on the backwards letter span task, $r(220) = -0.27$, $p < .01$, CI [−0.39, −0.14]. Also replicating what was observed in Study 1, there was no significant relation between the NDE generated in the comparison to a standard task and performance on the spatial index, $r(221) = -0.09$, $p = .18$, CI [−0.22, 0.04]. This remained the case even after controlling for performance on the backwards letter span task, $r(220) = -0.09$, $p = .20$, CI = [−0.21, 0.05].

Further, the correlation between the spatial index and the NDE of the simultaneous comparison task was significantly larger than the correlation between the spatial index and the NDE of the comparison to standard task, as indicated by the fact that the 95% CI of the difference between two correlations did not cross zero (Zou, 2007), 95% CI [0.06, 0.37].

The NDEs observed in the error data did not correlate with the spatial index; simultaneous comparison task, $r(221) = 0.06$, $p = .36$; comparison to standard, $r(221) = 0.11$, $p = .10$.

Discussion

Study 2 served as a preregistered replication of Study 1. Critically, we replicated our previous finding that the NDE elicited when comparing simultaneously presented numbers is more strongly negatively related to performance on the spatial index than is NDE elicited when comparing numbers to a standard. Interestingly, we also observed a significant correlation between the NDEs elicited by the simultaneous comparison task and the comparison to standard task. This contrasts with Study 1, where no such correlation was observed. This discrepancy may be attributed to the attenuation of the observed correlation given the relatively poorer reliability of the simultaneous comparison NDE in Study 1. As in Study 1, these data are consistent with the theory that the size of the NDE is related to one's ability to summon a more spatially precise representation of a spatially oriented mental number line. Importantly, this pattern was more strongly observed for the NDE elicited by the simultaneous comparison task compared to that elicited by the comparison to standard task, suggesting that the mechanism in which the two comparison tasks generate the NDE are independent. This aligns with the hypothesis of independent mechanisms for generating the NDE through the two tasks as evidenced by low between-task correlations (see Maloney et al., 2010) and differential interactions with working memory (see Maloney et al., 2019).

General Discussion

In two separate studies, one of which was a preregistered replication, we tested the hypothesis that the NDE that is generated during numerical comparison is negatively related to one's spatial competence. We generated our novel hypotheses based on the three dominant accounts of the NDE. According to two accounts of the NDE (the *representational overlap account* [Restle, 1970] and the *working memory account* [van Dijck & Fias, 2011]), the NDE arises due to the placement of numbers along an internal mental number line that is spatially represented. The *response competition account* (Verguts et al., 2005), on the other hand, does not posit a spatial mental number line in its explanation of the NDE. We investigated spatial competence using a heterogeneous approach, using specific spatial tasks, as well as a homogeneous one, using the spatial index. Consistent across two studies, we observed that the NDE that is generated in the simultaneous comparison task is more strongly related to performance on spatial tasks than is the NDE generated in the comparison to a standard task. The results of these two studies, while complex, provide potential insights into the role that spatial processing may play in numerical comparison.

Spatial Processing and Numerical Comparison

There are three dominant theories of the NDE. Both the *representational overlap account* (Restle, 1970) and the *working memory account* (van Dijck & Fias, 2011) posit a role for spatial representation and thus (in principle at least) spatial competency in the generation of the NDE. The *response competition account* (Verguts et al., 2005), on the other hand, does not. Importantly, the results of the present studies point to an interesting possibility wherein the theory that best accounts for the NDE depends on the comparison task used. Specifically, in line with our predictions drawn from the *representational overlap account* (Restle, 1970) and the *working memory account* (van Dijck & Fias, 2011), these accounts may better explain the NDE generated within the simultaneous comparison task, given its stronger correlation with performance on the spatial index, than can the *response competition account* (Verguts et al., 2005). Specifically, smaller NDEs (which are believed to be indicative of more spatially precise mental number line; De Smedt et al., 2009, 2013; Holloway & Ansari, 2009) are associated with better performance on the spatial

index, our measure of spatial competence. Conversely, the *response competition account* (Verguts et al., 2005) may better explain the NDE generated within the comparison to a standard task, given its weaker link to performance on the spatial index. Said another way, the current data are consistent with the theory that the comparison of two simultaneously presented numbers is more likely to be done relying on a spatially represented mental number line (Restle, 1970; van Dijck & Fias, 2011), whereas comparing one number to a standard is more likely to be done by comparing the relative association of each number with a "larger than" and "smaller than" response (Verguts et al., 2005). Importantly, given our current findings, the original psychophysical model proposed by Moyer and Landauer (1967) may also account for the NDE that arises within the comparison to a standard task as it does not assert a spatial property in its account.

Another interesting possibility is that participants rely on different comparison strategies on different trials and that the degree to which they rely on these different strategies varies as a function of the task set. One strategy may rely on a mental number line (as proposed by Restle, 1970 and van Dijck & Fias, 2011), whereas the other strategy may involve rely on the "larger" versus "smaller" information associated with a given number (as proposed by Verguts et al., 2005). Thus, it may well be the case that when comparing two simultaneously presented numbers, participants will make use of a mental number line more often than what they do when comparing one number to a standard. Thus, it may not be the case that one theory of the NDE is "correct" and the others are "incorrect" but rather the data are consistent with all three theories, to varying degrees and under varying conditions (i.e., varying formats of numerical comparison tasks). Below, we further integrate the current data into existing theories of differences that exist between the two comparison tasks.

Implications for Numerical Comparison Research

The apparent difference in the magnitude of relation observed here between the NDEs generated within the simultaneous comparison and comparison to standard tasks is consistent with Maloney et al.'s (2019) claim that the mechanism underlying the two NDEs are not the same (despite them being treated as equivalent in the literature). Maloney et al. (2019) suggested that when comparing two simultaneously presented numbers, the comparators are stored in WM as in the *working memory account* (van Dijck & Fias, 2011), whereas when comparing one number to a standard, the comparator is stored in LTM. Extending this account to the current work, when comparing simultaneously presented numbers, the comparators held in WM could be spatial in nature. In addition, when comparing a number to a standard, the present results can be interpreted to suggest that the comparison in LTM may not be spatial, or at least, is less spatial than a comparison that occurs in WM. From this perspective, the *working memory account* posited by van Dijck and Fias (2011) may best explain the NDE that arises when comparing two simultaneously presented numbers and the *response competition account* posited by Verguts et al. (2005) may best explain the NDE that arises when comparing one number to a standard. While one could suggest that the *representational overlap account* (Restle, 1970) could explain the NDE that arises in the simultaneous comparison task (as it does correlate strongly with performance on the spatial index), that theory is inconsistent with the findings by Maloney et al. (2019) who found that the NDE that arises in the Simultaneous Comparison Task was underadditive with a verbal WM load. Maloney et al. (2019) thus concluded that the comparison that occurs in the simultaneous comparison task is done using a representation that is stored in WM, not LTM, as is posited by the *representational overlap account* (Restle, 1970).

Split-Half Reliability of the NDEs

The data presented herein are consistent with the theory that comparing simultaneously presented numbers is more likely to be occurring on a spatially represented mental number line than is comparing a number to a standard. However, one issue that must be taken into consideration is that of the poor reliabilities of the NDEs in the comparison tasks. A possible explanation for the poor reliability observed in the present data is the low number of numerical comparison trials. Although the present study used a similar number of numerical comparison trials to that which has been used in previous individual difference research (e.g., Bugden & Ansari, 2011; Holloway & Ansari, 2008, 2009; Reynvoet et al., 2009; Sasanguie et al., 2017; Sekuler & Mierkiewicz, 1977), more recent work has shown that higher trial counts are needed to establish stronger split-half reliabilities when using cognitive tasks for individual difference research (Rouder et al., 2023). Thus, to obtain more reliable measures of the NDE, a greater number of trials than what was used here are needed (see Krajcsi & Szűcs, 2022).

Low reliability within a task is problematic as it can lead to a skewed estimate of the true correlation between performance on two tasks (Spearman, 1910). Within the current context, one could hypothesize that the reason the correlation between the simultaneous comparison NDE and performance on the spatial index is larger than that between the spatial index and the comparison to a standard task NDE is due to differing reliabilities of the NDE. However, the poor reliabilities do not appear to provide a viable explanation for the differential relations between the NDEs and performance on spatial tasks for two reasons. First, reliabilities were poor for both tasks. Thus, while the poor reliabilities make it difficult to estimate the true correlation, they have likely attenuated the true correlations to the same degree. Second, while the relative strength of the reliabilities changes across Studies 1 and 2, (i.e., in Study 1 the simultaneous comparison NDE is numerically less reliable, whereas in Study 2 it is numerically more reliable), the pattern of data remains the same within the two studies. That is, in both Studies 1 and 2, the correlation between the NDEs and the spatial index is larger for the simultaneous comparison tasks than for the comparison to a standard task. Thus, it is not the case that the comparison task with the numerically smaller reliability of the NDE is always the task with the lowest correlation with the spatial index. Thus, while the poor reliabilities should temper one's enthusiasm regarding the robustness of the reported results, it seems reasonable to conclude that the correlation between the NDE and performance on the spatial index is larger for the simultaneous comparison task than for the comparison to a standard task.

$$r_{XY} = R_{XY}\sqrt{\text{reliability}_X \times \text{reliability}_Y}. \quad (1)$$

Given that the observed correlations here are attenuated by the low reliabilities of the NDEs, they may not accurately reflect the strength of the correlation between the NDE within each numerical comparison task and spatial competence. One can correct for attenuation and assess the degree of potential attenuation using the

attenuation formula presented in Equation (1) where r_{XY} is the observed correlation between measures of two constructs, reliability$_X$ is the reliability of the measure of construct X, reliability$_Y$ is the reliability of measure of construct Y, and R_{XY} is the "true" correlation. If we assume that the true correlation between the NDE of each numerical comparison task and the spatial tasks are 1 and the reliability of the spatial tasks are also 1, then using Equation (1), one obtains a maximum corrected correlation of 0.39 and 0.55 for NDE in the simultaneous comparison task and the NDE in the comparison to standard task, respectively, in Study 1. Note that the observed correlations here were −0.31 and −.09, respectively, hence the observed correlation was strongest in the task with the lowest corrected correlation/reliability. In Study 2, one obtains a maximum corrected correlation of 0.66 and 0.50 for the NDE in the simultaneous comparison task and the NDE in the comparison to standard task, respectively. Given that the true correlation is unlikely to be 1 and that the reliabilities of spatial tasks are lower than 1 (e.g., reliabilities of spatial tasks are between 0.58 and 0.96 in Study 2), the degree of potential attenuation is likely to be smaller than the calculated maximum corrected values.

Critically, while low reliability attenuates the correlation, it is possible that the observed correlation is closer to the true correlation due spurious correlations between measurement error within each task (see Nimon et al., 2012). For example, Nimon et al. (2012) demonstrated that spurious correlations between the measurement error within each task or between measurement error and true scores can increase the correlation between two tasks, effectively counteracting the attenuating effect of low reliabilities. This may explain why the correlation between the NDE elicited in the simultaneous comparison task and spatial tasks is higher than the split-half reliability of the NDE elicited in the simultaneous comparison task. Importantly, it should be noted that this specific observation was not replicated in Experiment 2.

Nevertheless, the results of the present studies have important implications for all researchers using numerical comparison tasks. Indeed, given the poor reliabilities of the NDE, researchers examining relations between the NDE and performance on other measures or other individual differences should exercise caution, especially when interpreting null findings. At a minimum, researchers should assess the split-half reliabilities of their numerical comparison measures and take into consideration whether their findings, especially any null findings, may be due to poor reliability of their measures.

Generation of the NDEs, One Mechanism or Multiple?

One of the relatively novel theories that we have posited herein is the theory that the mechanism that gives rise to the NDE may differ as a function of task. While we cannot draw firm conclusions based on our data alone, when taken together with the broader literature, there is a growing body of evidence to support this argument. In terms of the present studies, the fact that the NDE that is generated in the simultaneous comparison task more strongly correlates with performance on the spatial index than does the NDE generated within the comparison task points to, at least, two separate mechanisms underlying these effects. One can also look at how highly the NDEs generated within the two comparison tasks correlate with one another. Paralleling that which was reported by Maloney et al. (2010), while there are small correlations between the two NDEs, these correlations are not as high as one would expect were the tasks measuring the exact same thing. Given the low reliabilities of the NDEs, the true between-task correlations are likely higher than what we are seeing here. That said, within the broader literature, recent work by Maloney et al. (2019) demonstrated that completing numerical comparison tasks while maintaining a verbal WM load causes the NDE within the simultaneous comparison task to decrease in size while the same WM load causes the NDE in comparison to a standard to increase in size. Thus, despite the expectation that the NDE generated within each task should be similarly affected by the same WM load, the observed differences suggest that the mechanisms underlying the NDE in the two tasks may be, to some extent, independent. That said, the two are unlikely to be completely distinct. For example, individual differences in numerical magnitude processing are argued to be linked to the size of the NDE (De Smedt et al., 2009; Holloway & Ansari, 2009). As such, individuals with stronger numerical magnitude processing skills may simply generate smaller NDE irrespective of the comparison task used (i.e., comparison strategy). Nonetheless, the current data and those reported by Maloney et al. (2010, 2019) constitute a growing body of literature that calls into question whether the NDEs that arise in the two symbolic numerical comparison tasks are generated due to the same underlying mechanisms.

Conclusion

In two studies, we tested the hypothesis that the NDE is related to performance on spatial tasks. An interesting pattern of results emerged such that the NDEs generated during simultaneous comparison relate to performance on spatial tasks more strongly than the NDEs generated during the comparison to a standard task. These data are consistent with the theory that numerical comparisons can, but may not always, be completed using a spatial representation or spatial processing. The degree to which the comparison related to spatial representation and processing may be dependent upon the type of numerical comparison task used. These data also add to the nascent literature highlighting the importance of not using the two numerical comparison tasks interchangeably as they are unlikely to be indexing the same underlying processes. All that being said, the present results also highlight the challenges present in investigating individual differences using behavioral metrics emerging from experimental manipulations of basic cognitive processes (see Krajcsi et al., 2024; Rouder et al., 2023). The reliabilities of the NDE in both are tasks were poor reducing our confidence in the correlations between measures.

Constraints in Generality

Upon careful consideration of the research on numerical representation and the numerical distance effect, we believe that the findings possess broad applicability. These results have implications that span across diverse groups and contexts, extending beyond the specific demographic tested in the present study. Indeed, the numerical distance effect has been proven to be a robust phenomenon. It has been observed in various age groups, spanning from childhood (e.g., Bugden & Ansari, 2011) to older adulthood (e.g., Norris et al., 2015), across different professions (Hohol et al., 2020), within populations with learning disabilities (e.g., Ashkenazi et al., 2009 for developmental dyscalculia), and even in distinct cultural settings (e.g., Quinlan et al., 2020 for Mandarin number

formats). However, given the novelty of our findings, it remains unclear whether the mechanisms we posit to generate the NDE are consistent across all samples. It is plausible that there are different mechanisms of generating the NDE that have yet to be explored in the present study. Further research involving larger and more diverse groups is warranted to shed light on this aspect.

References

Alloway, T. P., & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology*, *106*(1), 20–29. https://doi.org/10.1016/j.jecp.2009.11.003

Ashkenazi, S., Mark-Zigdon, N., & Henik, A. (2009). Numerical distance effect in developmental dyscalculia. *Cognitive Development*, *24*(4), 387–400. https://doi.org/10.1016/j.cogdev.2009.09.006

Bugden, S., & Ansari, D. (2011). Individual differences in children's mathematical competence are related to the intentional but not automatic processing of Arabic numerals. *Cognition*, *118*(1), 32–44. https://doi.org/10.1016/j.cognition.2010.09.005

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press. https://doi.org/10.1017/CBO9780511571312

Castronovo, J., & Gobel, S. M. (2012). Impact of high mathematics education on the number sense. *PLOS ONE*, *7*(4), Article e33832. https://doi.org/10.1371/journal.pone.0033832

Daker, R. J., Delage, V., Maloney, E. A., & Lyons, I. M. (2022). Testing the specificity of links between anxiety and performance within mathematics and spatial reasoning. *Annals of the New York Academy of Sciences*, *1512*(1), 174–191. https://doi.org/10.1111/nyas.14761

De Smedt, B., Noël, M.-P., Gilmore, C., & Ansari, D. (2013). How do symbolic and non-symbolic numerical magnitude processing skills relate to individual differences in children's mathematical skills? A review of evidence from brain and behavior. *Trends in Neuroscience and Education*, *2*(2), 48–55. https://doi.org/10.1016/j.tine.2013.06.001

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

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General*, *122*(3), 371–396. https://doi.org/10.1037/0096-3445.122.3.371

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*(3), 626–641. https://doi.org/10.1037/0096-1523.16.3.626

Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PLOS ONE*, *10*(3), Article e0121945. https://doi.org/10.1371/journal.pone.0121945

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Ferguson, A. M., Maloney, E. A., Fugelsang, J., & Risko, E. F. (2015). On the relation between math and spatial ability: The case of math anxiety. *Learning and Individual Differences*, *39*, 1–12. https://doi.org/10.1016/j.lindif.2015.02.007

Festinger, L. (1943). Studies in decision: I. Decision-time, relative frequency of judgment and subjective confidence as related to physical stimulus difference. *Journal of Experimental Psychology*, *32*(4), 291–306. https://doi.org/10.1037/h0056685

Gevers, W., Verguts, T., Reynvoet, B., Caessens, B., & Fias, W. (2006). Numbers and space: A computational model of the SNARC effect. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(1), 32–44. https://doi.org/10.1037/0096-1523.32.1.32

Göbel, S. M., Shaki, S., & Fischer, M. H. (2011). The cultural number line: A review of cultural and linguistic influences on the development of number processing. *Journal of Cross-Cultural Psychology*, *42*(4), 543–565. https://doi.org/10.1177/0022022111406251

Hegarty, M., & Waller, D. (2004). A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence*, *32*(2), 175–191. https://doi.org/10.1016/j.intell.2003.12.001

Herrera, A., Macizo, P., & Semenza, C. (2008). The role of working memory in the association between number magnitude and space. *Acta Psychologica*, *128*(2), 225–237. https://doi.org/10.1016/j.actpsy.2008.01.002

Höffler, T. N. (2010). Spatial ability: Its influence on learning with visualizations—A meta-analytic review. *Educational Psychology Review*, *22*(3), 245–269. https://doi.org/10.1007/s10648-010-9126-7

Hohol, M., Willmes, K., Nęcka, E., Brożek, B., Nuerk, H.-C., & Cipora, K. (2020). Professional mathematicians do not differ from others in the symbolic numerical distance and size effects. *Scientific Reports*, *10*(1), Article 11531. https://doi.org/10.1038/s41598-020-68202-z

Holloway, I. D., & Ansari, D. (2008). Domain-specific and domain-general changes in children's development of number comparison. *Developmental Science*, *11*(5), 644–649. https://doi.org/10.1111/j.1467-7687.2008.00712.x

Holloway, I. D., & Ansari, D. (2009). Mapping numerical magnitudes onto symbols: The numerical distance effect and individual differences in children's mathematics achievement. *Journal of Experimental Child Psychology*, *103*(1), 17–29. https://doi.org/10.1016/j.jecp.2008.04.001

Krajcsi, A., Chesney, D., Cipora, K., Coolen, I., Gilmore, C., Inglis, M., Libertus, M., Nuerk, H.-C., Simms, V., & Reynvoet, B. (2024). Measuring the acuity of the approximate number system in young children. *Developmental Review*, *72*, Article 101131. https://doi.org/10.1016/j.dr.2024.101131

Krajcsi, A., Lengyel, G., & Kojouharova, P. (2016). The source of the symbolic numerical distance and size effects. *Frontiers in Psychology*, *7*, Article 1795. https://doi.org/10.3389/fpsyg.2016.01795

Krajcsi, A., & Szűcs, T. (2022). Symbolic number comparison and number priming do not rely on the same mechanism. *Psychonomic Bulletin & Review*, *29*(5), 1969–1977. https://doi.org/10.3758/s13423-022-02108-x

Linn, M. C., & Petersen, A. C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development*, *56*(6), 1479–1498. https://doi.org/10.2307/1130467

Maloney, E. A., Risko, E. F., O'Malley, S., & Besner, D. (2009). Tracking the transition from sublexical to lexical processing: On the creation of orthographic and phonological lexical representations. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *62*(5), 858–867. https://doi.org/10.1080/17470210802578385

Maloney, E. A., Barr, N., Risko, E. F., & Fugelsang, J. A. (2019). Verbal working memory load dissociates common indices of the numerical distance effect: Implications for the study of numerical cognition. *Journal of Numerical Cognition*, *5*(3), 337–357. https://doi.org/10.5964/jnc.v5i3.155

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. https://doi.org/10.1016/j.actpsy.2010.01.006

Manna, C. B. G., Tenke, C. E., Gates, N. A., Kayser, J., Borod, J. C., Stewart, J. W., McGrath, P. J., & Bruder, G. E. (2010). EEG hemispheric asymmetries during cognitive tasks in depressed patients with high versus low trait anxiety. *Clinical EEG and Neuroscience*, *41*(4), 196–202. https://doi.org/10.1177/155005941004100406

Meyerhoff, H. S., Moeller, K., Debus, K., & Nuerk, H.-C. (2012). Multi-digit number processing beyond the two-digit number range: A combination of sequential and parallel processes. *Acta Psychologica*, *140*(1), 81–90. https://doi.org/10.1016/j.actpsy.2011.11.005

Mix, K. S., & Cheng, Y.-L. (2012). The relation between space and math: Developmental and educational implications. *Advances in Child Development and Behavior*, *42*, 197–243. https://doi.org/10.1016/B978-0-12-394388-0.00006-X

Mix, K. S., Hambrick, D. Z., Satyam, V. R., Burgoyne, A. P., & Levine, S. C. (2018). The latent structure of spatial skill: A test of the 2 × 2 typology. *Cognition*, *180*, 268–278. https://doi.org/10.1016/j.cognition.2018.07.012

Money, J., Walker, H. T., Jr., & Duane, A. (1965). Development of direction sense and three syndromes of impairment. *Slow Learning Child*, *11*(3), 145–155. https://doi.org/10.1080/0156655650110304

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature*, *215*(5109), 1519–1520. https://doi.org/10.1038/2151519a0

Nimon, K., Zientek, L. R., & Henson, R. K. (2012). The assumption of a reliable instrument and other pitfalls to avoid when considering the reliability of data. *Frontiers in Psychology*, *3*, Article 102. https://doi.org/10.3389/fpsyg.2012.00102

Norris, J. E., McGeown, W. J., Guerrini, C., & Castronovo, J. (2015). Aging and the number sense: Preserved basic non-symbolic numerical processing and enhanced basic symbolic processing. *Frontiers in Psychology*, *6*, Article 999. https://doi.org/10.3389/fpsyg.2015.00999

Nuerk, H.-C., Moeller, K., & Willmes, K. (2015). Multi-digit number processing: Overview, conceptual clarifications, and language influences. In R. Cohen Kadosh & A. Dowker (Eds.), *The Oxford handbook of numerical cognition* (pp. 106–139). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199642342.013.021

Nuerk, H.-C., Weger, U., & Willmes, K. (2001). Decade breaks in the mental number line? Putting the tens and units back in different bins. *Cognition*, *82*(1), B25–B33. https://doi.org/10.1016/S0010-0277(01)00142-1

Nuerk, H.-C., Weger, U., & Willmes, K. (2005). Language effects in magnitude comparison: Small, but not irrelevant. *Brain and Language*, *92*(3), 262–277. https://doi.org/10.1016/j.bandl.2004.06.107

Pronk, T., Molenaar, D., Wiers, R. W., & Murre, J. (2022). Methods to split cognitive task data for estimating split-half reliability: A comprehensive review and systematic assessment. *Psychonomic Bulletin & Review*, *29*(1), 44–54. https://doi.org/10.3758/s13423-021-01948-3

Quinlan, P. T., Cohen, D. J., & Liu, X. (2020). Further insights into the operation of the Chinese number system: Competing effects of Arabic and Mandarin number formats. *Memory & Cognition*, *48*(8), 1472–1483. https://doi.org/10.3758/s13421-020-01065-x

Restle, F. (1970). Speed of adding and comparing numbers. *Journal of Experimental Psychology*, *83*(2, Pt 1), 274–278. https://doi.org/10.1037/h0028573

Retanal, F., Delage, V., Maloney, E., & Risko, E. (2022). *Numerical comparison and Spatial abilities*. Open Science Framework. https://doi.org/10.17605/OSF.IO/FZX3Y

Reynvoet, B., De Smedt, B., & Van den Bussche, E. (2009). Children's representation of symbolic magnitude: The development of the priming distance effect. *Journal of Experimental Child Psychology*, *103*(4), 480–489. https://doi.org/10.1016/j.jecp.2009.01.007

Rouder, J. N., Kumar, A., & Haaf, J. M. (2023). Why many studies of individual differences with inhibition tasks may not localize correlations. *Psychonomic Bulletin & Review*, *30*(6), 2049–2066. https://doi.org/10.3758/s13423-023-02293-3

Sasanguie, D., Lyons, I. M., De Smedt, B., & Reynvoet, B. (2017). Unpacking symbolic number comparison and its relation with arithmetic in adults. *Cognition*, *165*, 26–38. https://doi.org/10.1016/j.cognition.2017.04.007

Sekuler, R., & Mierkiewicz, D. (1977). Children's judgments of numerical inequality. *Child Development*, *48*(2), 630–633. https://doi.org/10.2307/1128664

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701–703. https://doi.org/10.1126/science.171.3972.701

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of PsYchology*, *3*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., & Newcombe, N. S. (2013). The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin*, *139*(2), 352–402. https://doi.org/10.1037/a0028446

van Dijck, J.-P., & Fias, W. (2011). A working memory account for spatial–numerical associations. *Cognition*, *119*(1), 114–119. https://doi.org/10.1016/j.cognition.2010.12.013

Verguts, T., Fias, W., & Stevens, M. (2005). A model of exact small-number representation. *Psychonomic Bulletin & Review*, *12*(1), 66–80. https://doi.org/10.3758/BF03196349

Viarouge, A., Hubbard, E. M., & McCandliss, B. D. (2014). The cognitive mechanisms of the SNARC effect: An individual differences approach. *PLOS ONE*, *9*(4), Article e95756. https://doi.org/10.1371/journal.pone.0095756

Wechsler, D. (1997). *WAIS-III administration and scoring manual*. The Psychological Corporation.

Yan, T. (2008). Nondifferentiation. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methodology* (p. 521). Sage Publications. https://doi.org/10.4135/9781412963947

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods*, *12*(4), 399–413. https://doi.org/10.1037/1082-989X.12.4.399

Received February 28, 2023
Revision received June 16, 2024
Accepted June 22, 2024 ■