Psychological Research (2021) 85:1248–1271
https://doi.org/10.1007/s00426-020-01299-z

ORIGINAL ARTICLE

Probing the mechanisms underlying numerosity-to-numeral mappings and their relation to math competence

Darren J. Yeo[1,2] · **Gavin R. Price**[1]

Received: 18 April 2019 / Accepted: 28 January 2020 / Published online: 14 February 2020

Abstract

Numerosity estimation performance (e.g., how accurate, consistent, or proportionally spaced (linear) numerosity-numeral mappings are) has previously been associated with math competence. However, the specific mechanisms that underlie such a relation is unknown. One possible mechanism is the mapping process between numerical sets and symbolic numbers (e.g., Arabic numerals). The current study examined two hypothesized mechanisms of numerosity-numeral mappings (item-based "associative" and holistic "structural" mapping) and their roles in the estimation-and-math relation. Specifically, mappings for small numbers (e.g., 1–10) are thought to be associative and resistant to calibration (e.g., feedback on accuracy of estimates), whereas holistic "structural" mapping for larger numbers (e.g., beyond 10) may be supported by flexibly aligning a numeral "response grid" (akin to a ruler) to an analog "mental number line" upon calibration. In 57 adults, we used pre- and post-calibration estimates to measure the range of continuous associative mappings among small numbers (e.g., a base range of associative mappings from 1 to 10), and obtained measures of math competence and delayed multiple-choice strategy reports. Consistent with previous research, uncalibrated estimation performance correlated with calculation competence, controlling for reading fluency and working memory. However, having a higher base range of associative mappings was not related to estimation performance or any math competence measures. Critically, discontinuity in calibration effects was typical at the individual level, which calls into question the nature of "holistic structural mapping". A parsimonious explanation to integrate previous and current findings is that estimation performance is likely optimized by dynamically constructing numerosity-numeral mappings through the use of multiple strategies from trial to trial.

Introduction

Estimating the number of items in a set (i.e., numerical estimation) is an efficient alternative to counting and its performance is thought by many to either directly reflect the quality of the mappings between the encoded numerosity of a set (e.g., dot arrays) and symbolic estimates (i.e., verbal and Arabic numerals, such as "five" and "5"), or indirectly reflect the acuity of the encoded representations of the numerosities themselves (e.g., Brankaer, Ghesquière, & De Smedt, 2014; Ebersbach & Erz, 2014; Izard & Dehaene, 2008; Jang & Cho, 2018; Libertus, Feigenson, Halberda, & Landau, 2014; Libertus, Odic, Feigenson, & Halberda, 2016; Lipton & Spelke, 2005; Mundy & Gilmore, 2009). Furthermore, the quality of those mappings is suggested to relate to math competence. Specifically, individuals who make more accurate, consistent, or proportionally spaced (i.e., linear) estimates tend to demonstrate higher math competence (Alvarez et al., 2017; Bartelet, Vaessen, Blomert, & Ansari, 2014; Booth & Siegler, 2006; Castronovo & Göbel, 2012; Chesney, Bjalkebring, & Peters, 2015; Guillaume, Gevers, & Content, 2016; Libertus et al., 2016; Lyons, Price, Vaessen, Blomert, & Ansari, 2014; Mazzocco, Feigenson, & Halberda, 2011a; Mejias, Grégoire, & Noël, 2012; Mejias, Mussolin, Rousselle, Grégoire, & Noël, 2012; Mejias & Schiltz, 2013; Pinheiro-Chagas et al., 2014; Wong, Ho, & Tang, 2016a, b). However, the acuity of representations of numerosity and the quality of numeral-numerosity mappings are

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00426-020-01299-z) contains supplementary material, which is available to authorized users.

✉ Gavin R. Price
gavin.price@vanderbilt.edu

[1] Department of Psychology and Human Development, Peabody College, Vanderbilt University, 230 Appleton Place, Nashville, TN 37203, USA

[2] Division of Psychology, School of Social Sciences, Nanyang Technological University, 48 Nanyang Avenue, Singapore 639818, Singapore

not the only plausible explanations for the relation between estimation performance and math outcomes. The mapping processes between perceptual representations of numerosity and symbolic estimates may also underlie the relation. For example, some numerosity-numeral mappings may have been acquired through item-to-item associations (Sullivan & Barner, 2013), and the direct retrieval of these acquired associations between numerosities and symbolic numbers may be critical for both estimation and math. Alternatively, most mappings between a numeral and an unfamiliar numerosity are not stable (Izard & Dehaene, 2008), and their malleability during estimation tasks may be supported by similar ordinal relations in the discrete structure of symbolic numbers (e.g., "five" comes after "two") and the analog structure of representations of numerosity (e.g., {●●●●●} > {●●}) (for a review, see Carey & Barner, 2019; Izard & Dehaene, 2008; Sullivan & Barner, 2013). This holistic structural mapping that is thought to underlie most numerosity-numeral mappings may also be important for both estimation and math. Therefore, at present, the cognitive mechanisms that underlie the relation between estimation performance and math competence remain unknown. The current study addresses this gap by examining the roles of two hypothesized numerosity-numeral mapping mechanisms—item-based associative mapping and holistic structural mapping—in the estimation–math relation.

Associative mappings versus structural mappings as a function of set size

Mechanisms underlying numerosity-numeral mappings during numerical estimation have primarily been investigated using external (i.e., experimenter-provided) calibration paradigms. In such paradigms, participants are typically first asked to make a series of spontaneous estimates of sets of objects (i.e., uncalibrated). Thereafter, participants are provided with opportunities to calibrate their estimates against an external reference before repeating or continuing with the rest of the task. They may be shown a visual reference (e.g., "there are *n* dots in this array") (e.g., Krueger, 1984), given feedback after an estimate has been made (e.g., "there were actually *n* dots in the previous array") (e.g., Price, Clement, & Wright, 2014), or provided with an upper bound without a visual reference (e.g., "the largest set of dots you will see is *n* dots") (e.g., Sullivan & Barner, 2013). The change in participants' distribution of estimates for each tested numerosity before and after calibration (Sullivan & Barner, 2013, 2014), or between two calibration conditions (Izard & Dehaene, 2008), is assessed and the susceptibility of the numerosity-numeral mappings to calibration (hereafter, "calibration effects") is used to make inferences about the underlying mechanisms. Based on such research, two distinct mapping mechanisms are suggested to co-exist—item-based "associative mapping" and holistic "structural mapping" (for a review, see Carey & Barner, 2019; Sullivan & Barner, 2013, 2014).

Associative mappings are thought to be independent item-specific associations between particular numerals (e.g., "five" or "5") and mental representations of numerosity (e.g., fiveness in 5 crackers, 5 people, etc.) (Sullivan & Barner, 2013, 2014). However, such mappings may not be perfectly accurate and/or consistent due to the supposedly approximate nature of our mental representations of numerosity (Dehaene, 2007; Izard & Dehaene, 2008). A signature of strong associative mappings is that they are resistant to external calibration, presumably because they involve a direct retrieval of the item-specific associations (Sullivan & Barner, 2013, 2014). These associative mappings are thought to result from accumulated experience with pairings between certain numerals and perceptual representations of numerosity (Dehaene & Mehler, 1992; Lipton & Spelke, 2005; Verguts & Fias, 2004), and are more likely to support the estimation of small numerosities than of larger ones (Sullivan & Barner, 2013, 2014). Such associative mappings are commonly observed to be a continuous extension of the subitizing range of 1 through about 4 (e.g., one through nine) (see Fig. 1). Hereafter, we refer to this continuous set of associative mappings the "base range of associative mappings". It remains unclear whether true associative mappings for much larger numerosities (e.g., 50 or 100) are logically plausible. Consistent with this experience-dependent account, 5–7 year-olds have a smaller base range of associative mappings (up to about 6) (Sullivan & Barner, 2014) than adults (up to about 12) (Sullivan & Barner, 2013).

As it is logically impossible for an individual to have associative experience for every numeral-numerosity mapping, especially for large numbers (e.g., one thousand), the majority of numeral-numerosity mappings are thought to be inferred from those that are associatively experienced (Alvarez et al., 2017; Le Corre & Carey, 2007; Sullivan & Barner, 2013, 2014). Specifically, it has been proposed that the associative mappings involving small numbers, no matter how few, are essential for supporting the holistic linking between an analog mental number line and the discrete system of numerals on the basis of their analogous ordinal structures (i.e., the numerosity/numeral that comes after is greater), which in turn enables the inferential processes for larger numbers without associative experiences (Carey & Barner, 2019; Carey, Shusterman, Haward, & Distefano, 2017; Le Corre & Carey, 2007). These inferred mappings are referred to as structural mappings and are thought to be causally interdependent, such that the whole range of mappings can be influenced simultaneously by calibration even with a single-value feedback (Carey & Barner, 2019; Izard & Dehaene, 2008; Krueger, 1984; Minturn & Reese, 1951; Sullivan & Barner, 2013, 2014). For instance, when

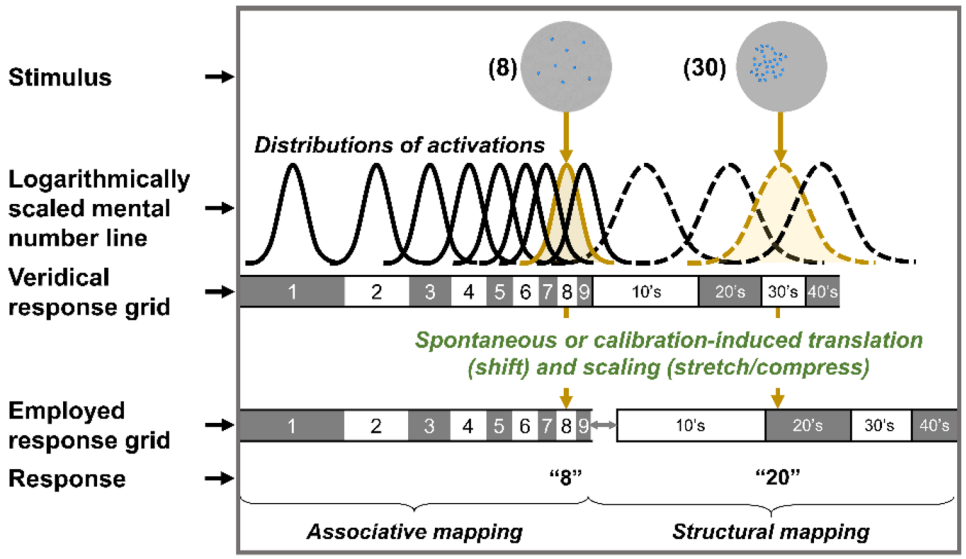

Fig. 1 A schematic summary of Izard and Dehaene's (2008) response-grid model, and the associative mapping and structural mapping mechanisms proposed by Sullivan and Barner (2013) involved in numerical estimation in a hypothetical individual. Individuals typically spontaneously underestimate large quantities (e.g., assigning "20" to an array of 30 dots) (Izard & Dehaene, 2008), because the spontaneously employed response grid is typically not aligned accurately with the mental number line. An external calibration is thought to transform the employed response grid with the goal that it could be more like the veridical response grid

participants are provided with a particular numeral-numerosity mapping visually (e.g., an array of 30 dots as "30"), they tend not to only modify their subsequent estimates whenever they perceive 30 dots according to the learned mapping, but also estimates for all other numerosities (Izard & Dehaene, 2008).

How does such interdependency of mappings occur? Izard and Dehaene (2008) proposed that the analog number system or mental number line is divided into segments, and each segment is associated with a different numeral label (i.e., a "response grid", which serves as an interface between the analog and discrete systems) (see Fig. 1). The response grid that is actually employed is typically not accurately aligned with the mental number line, with the exception of the base range of associative mappings. In other words, much of the "employed response grid" typically differs from the "veridical response grid". However, the employed response grid can be readily transformed by means of a translation and/or a scaling parameter(s) with an internal (spontaneous) or external calibration (Izard & Dehaene, 2008). The employed response grid is also thought to remain stable throughout the course of a task requiring a series of estimates (Izard & Dehaene, 2008).

Hence, this response-grid model explicitly describes how the analog mental number line and the symbolic numeral system are holistically mapped, which in turn gives rise to the observed interdependency between numerosity-numeral mappings.

Once the numerosity is encoded, an association between a numerosity and a symbolic response is thought to be retrieved from a stably aligned, employed response grid (Izard & Dehaene, 2008). If numerosities are processed instantaneously and in parallel, as commonly assumed in most computational models of numerosity encoding (Dehaene, 2007; Dehaene & Changeux, 1993; Stoianov & Zorzi, 2012; Verguts & Fias, 2004), it is conceivable that the most efficient strategy for labeling larger numerosities is to retrieve a numeral-numerosity mapping from the response grid (e.g., it looked like there are about 20 or a little less, so I labeled it as "18"; i.e., "benchmarking" strategy) instead of enumerating serially (Gandini, Ardiale, & Lemaire, 2010; Gandini, Lemaire, & Dufau, 2008). Hence, a key prediction pertaining to the effects of calibration is that if a set of 20 is calibrated to be assigned to "30", then numerals to be assigned to sets of 18, 19, 21, or 22 should be adjusted accordingly, especially if they are processed instantaneously and in parallel as approximately 20 (e.g., Piazza, Pinel, Le Bihan, & Dehaene, 2007). In other words, the calibration effects ought to be continuous across the range tested. In sum, associative and structural mappings are recruited during estimation as a function of set size, and the response-grid model strongly suggests that once mappings transition from associative to structural at a given numerosity, all subsequent numerosities should show a continuous set of structural mappings. Hence, a clear dissociation between a base range of associative mappings (resistant to calibration) and

continuous structural mappings (susceptible to calibration) is predicted. Such a set size-dependent dissociation holds true at the group level (Sullivan & Barner, 2013, 2014), although it may not at the individual level (Yeo, Wilkey, & Price, 2019). We found that 82% of 71 adults showed discontinuity in the effect of calibration across numerosities 8 through 350 (Yeo et al., 2019). However, because the numerosities were sparsely sampled in that previous study (i.e., 8, 12, 20, 35, 60, …, 350), there is a possibility that evidence for continuous structural mappings may be more apparent with a fine-grained sampling of a fairly large set of adjacent numbers (e.g., every number from 5 to 35).

Role of associative mappings in estimation and math competence

There is convergent evidence, across ages and numerical ranges tested, that higher accuracy (Bartelet et al., 2014; Booth & Siegler, 2006; Castronovo & Göbel, 2012; Guillaume et al., 2016; Lyons et al., 2014; Mejias, Grégoire, et al., 2012; Mejias, Mussolin, et al., 2012; Mejias & Schiltz, 2013), consistency (Guillaume et al., 2016; Libertus et al., 2016; Mazzocco et al., 2011a; Mejias, Grégoire, et al., 2012; Mejias, Mussolin, et al., 2012; Pinheiro-Chagas et al., 2014) and linearity (Alvarez et al., 2017; Booth & Siegler, 2006; Chesney et al., 2015; Wong et al., 2016a, b) of estimates are all related to higher math competence. These measures are thought to reflect the overall quality of numeral-numerosity mappings, but it remains unknown what cognitive mechanisms support the quality of the mappings.

Here, we focus on assertions that associative mappings are cognitively and behaviorally important for structural mappings, and, by extension, estimation performance in general. For instance, Carey et al. (2017) claim that structural mappings cannot be acquired "without having mapped at least some small numbers to [analog number system] values" (p. 254). Hence, a higher base range of associative mappings may support a more effective initial alignment of a response grid and the analog number system, and the subsequent transformation of the response grid, such that it can be more accurately aligned with the analog number system. Alternatively, a higher range may also strengthen the integrity of the link between the two systems, such that they are less likely to be perturbed, thereby resulting in more accurate, consistent, and linear mappings. If the analog mental number line is indeed harnessed for manipulation of symbolic number representations during math, as is commonly thought (e.g., Dehaene, Spelke, Pinel, Stanescu, & Tsivkin, 1999; Mundy & Gilmore, 2009; Pinheiro-Chagas, Dotan, Piazza, & Dehaene, 2017; Stoianov, 2014), the higher integrity of the mapping between the systems will enable more efficient and effective access and manipulation of the mental number line from numerals. This may account for why higher range of associative mappings may relate to both estimation performance and math competence. Another reason is that the range of the associative mappings may reflect higher acuity of the analog number representations, which have been proposed as a critical foundation for math competence (Mazzocco, Feigenson, & Halberda, 2011b; Starr, Libertus, & Brannon, 2013). On the other hand, higher math competence may also refine the acuity of mental number representations (Lyons, Bugden, Zheng, De Jesus, & Ansari, 2018; Mussolin, Nys, Content, & Leybaert, 2014; Suárez-Pellicioni & Booth, 2018), which in turn may increase the strength and extent of associative mappings. In either case, a higher range of associative mappings may reflect more precise representations of numerical magnitudes.

Furthermore, a higher range of associative mappings may form the foundation for improved structural mappings on an ad-hoc basis, as opposed to holistically. It has been proposed that structural mappings are supported by analogical reasoning (e.g., ::: is to ::::: as 6 is to…?) (Alvarez et al., 2017). However, the analogy could either be made only once holistically (i.e., "a single analogical mapping between the structure of the verbal count list as a whole and a range of corresponding [analog number system] values", Carey & Barner, 2019, p. 4), or on an ad-hoc basis (i.e., with individual analogous pairs, such as using sets of 5 to infer sets with multiples of 5). For instance, an individual with strong associative mapping only for five may be able to make analogical comparisons using five, but less effective in doing so from six or seven; an individual with associative mappings for five through ten may be better able to make a richer set of analogical comparisons, resulting in more proportionally spaced estimates across an extended range. Hence, although structural mappings are the predominant type, they are necessarily grounded in associative mappings insofar as base associations are a prerequisite for analogical inference. In sum, even with an analogy-based account that describes "structural mappings" in a more ad-hoc manner rather than a holistic manner, it is possible that a higher range of associative mappings may be related to better estimation performance and math competence.

In contrast, Sullivan and Barner (2014) argue that the role of associative mappings is fundamental but negligible given their scarcity even in adults (Sullivan & Barner, 2013), and suggest that it is "unlikely that those who are better at estimating (and thus better at math) have a relatively richer set of [associative mappings]" (p. 1753). Nonetheless, Sullivan and Barner (2013, 2014) did not report direct measures of overall estimation performance and math competence to support that conclusion. Hence, the role of associative mappings in estimation-and-math competencies remains unknown.

Current study

The aims of the current study were twofold. First, we aimed to examine whether individual differences in the base range of associative mappings are related to estimation performance and math competence, and whether they underlie previously observed relations between estimation performance and math competence, while controlling for potential confounding factors (fluency of retrieving symbol–referent associations and working memory). Hence, this study went beyond merely replicating a relation between math competence and a collection of commonly used indices of estimation performance (i.e., accuracy, consistency, and linearity of estimates), and examined whether the base range of associative mappings could underlie such relations. We hypothesized that a higher base range of associative mappings will be related to estimation performance as well as to math competence.

In previous studies, the onset of the range of structural mappings has been defined as the smallest numerosity that was influenced by calibration; correspondingly, the upper bound of the base range of associative mappings is defined as the largest numerosity that remained unaffected by calibration (Alvarez et al., 2017; Sullivan & Barner, 2013, 2014; Yeo et al., 2019). However, in those studies, numerosities were sparsely sampled across a wide range (e.g., 8, 12, 20 … 350) and individual differences in the associative mapping range could not be measured precisely. There is also a strong assumption that a participant who showed no calibration effects for 8 and 12 could not have shown calibration effects for 9 through 11. In other words, the effects of calibration ought to be continuous. Moreover, the calibration protocols used in the majority of previous studies (Alvarez et al., 2017; Sullivan & Barner, 2013, 2014; Yeo et al., 2019) were implicit (i.e., "the largest set of dots you will see is *n* dots") leaving it unclear as to how participants actually made use of the instruction to calibrate their subsequent estimates (i.e., whether they formed an association between *n* and their memory of the largest set which they saw during the uncalibrated condition, or whether they merely artificially restricted their estimates to within *n*). To address these concerns, we employed a modified explicit calibration protocol and sampled every numerosity within a narrow range in an estimation task with uncalibrated and calibrated conditions. The paradigm adopted is thus highly similar to that used by Izard and Dehaene (2008) in the development of the response-grid model, which not only allowed us to rule out the possibility that paradigm differences could explain any evidence of discontinuity in calibration effects, but also allowed us to measure the base range of associative mappings precisely in each individual.

Second, by assessing the continuity of calibration effects with a finer sampling resolution, and obtaining delayed multiple-choice strategy reports, we also aimed to extend our previous finding that estimation performance may not primarily reflect the use of a single holistic structural mapping via a response grid. We hypothesized that, if the response grid is essential for numerosity estimation, and that associative mappings and holistic structural mappings are clearly a function of set size, (1) calibration effects should be continuous after the onset of structural mapping, especially when numerosities are sampled without gaps, and (2) benchmarking (directly retrieving numerosity-numeral mappings from a response grid upon an instantaneous and parallel enumeration of items in a set) should be part of the strategy repertoire reported by most participants.

Methods

Participants

Fifty-seven undergraduate students (35 females; age range: 18.42–22.33 years, $M = 19.76$, $SD = 0.99$) participated in the study for course credit. The experimental protocol was approved by our university's Institutional Review Board and all participants provided written informed consent. Data from four additional participants were excluded due to incomplete or invalid data as a result of technical or experimenter error.

Procedure

The experiment was conducted one participant at a time, in a single session in a quiet room. All participants completed the tasks in the same order to minimize inter-individual differences in performance that could stem from variation in task order: uncalibrated condition followed by the calibrated condition of the estimation task, online questionnaire, standardized reading and math achievement tests, and working memory tasks. Uncalibrated estimation was necessarily administered prior to the calibrated condition so as to capture true spontaneous estimation performance. The stimuli for the estimation and working memory tasks were presented using E-Prime 2.0 (Psychology Software Tools, Pittsburgh, PA, USA) on a 21.5″ monitor that subtends a 43.3° × 25.2° visual angle with an approximate viewing distance of 60 cm. All participants were debriefed about the aims and predictions of the study at the end of the session. Due to scheduling issues, one participant completed the working memory tasks in a separate session 5 days after all the other tasks were completed.

Estimation task

The estimation task comprises three phases—a pre-calibration phase (hereafter, "uncalibrated"), a calibration manipulation phase, and a post-calibration phase (hereafter, "calibrated").

Uncalibrated condition Participants saw a series of blue-dot arrays presented at the center of a grey circular background (diameter of 23 cm, which covered a visual angle of 21.7°×21.7°; see Fig. 1 for example stimuli) against a black screen. On each test trial, the dot array was presented for 500 ms followed by a circular grey mask that prompted a response. The short presentation duration was chosen to prevent participants from serial counting of individual dots. Participants were given no information about the range of numerosities which they would see and were instructed to estimate the number of dots and enter their estimates using the numeric keypad on a computer keyboard as quickly and as accurately as possible. They were allowed to amend their estimates using the backspace key. After the entered response was confirmed by pressing the spacebar key, a central fixation cross within a circular grey background appeared for 1500 ms followed by the next set of dots. There were no practice trials and no feedback was given throughout the task.

Thirty-one numerosities from 5 through 35 were used and each numerosity was presented ten times, resulting in a total of 310 trials. The lower bound of five was chosen based on the assumption that quantities in the subitizing range (a mean of about 4 in typical adults, Kaufman, Lord, Reese, & Volkmann, 1949; Piazza, Fumarola, Chinello, & Melcher, 2011; Revkin, Piazza, Izard, Cohen, & Dehaene, 2008; Trick & Pylyshyn, 1994) are enumerated through a different process than numbers outside that range (Burr, Turi, & Anobile, 2010; Cutini, Scatturin, Basso Moro, & Zorzi, 2014; Hyde, 2011; Piazza et al., 2011; Pincham & Szucs, 2012; Revkin et al., 2008), and are, therefore, highly unlikely to be influenced by calibration. In other words, we assumed that numerosities one through four are associatively mapped and should not be influenced by calibration. The upper bound of 35 was based on the mode (8), mean (36), and median (20) of the smallest numerosity that was susceptible to calibration across 71 adults in a previous study undertaken in our lab (Yeo et al., 2019). To avoid any random clustering of similar numerosities, we divided the 310 trials into five implicit phases of 62 trials (i.e., each phase cycles through all 31 numerosities in a random order with each numerosity presented twice). All trials were completed within a single block with no breaks to maximize the continuity of estimates. The dots within each array were of the same size. To minimize the use of non-numerical visual cues such as occupied area and dot size, each numerosity was matched with every other numerosity on dot size and array density for half the trials (total occupied area and total luminance increase with numerosity), and on total occupied area for the other half (dot size decreases with numerosity and array density increases with numerosity) (Dehaene, Izard, & Piazza, 2005). This task was self-paced and took about 20–30 min.

Calibrated condition The calibrated condition had two phases, a calibration phase followed by a test phase. During the calibration phase, participants were shown arrays consisting of 45, 60, or 75 dots, with each numerosity presented four times (half of the trials were matched with those used in the test phase and in the uncalibrated estimation task in terms of dot size, and the other half were matched in terms of total occupied area) and were given feedback immediately after they made an estimate. Each array was presented for 2 s and the feedback was presented for 3 s. This longer presentation time relative to that in the test trials was to facilitate learning of the associations between the numerosity and the subsequent feedback. To maximize the calibration effects, such that they are large enough to distinguish associative mappings from structural mappings, we employed non-veridical feedback meant to induce an over-estimation (Alvarez et al., 2017; Izard & Dehaene, 2008; Sullivan & Barner, 2013, 2014). Through pilot testing, we established that a factor of 4/3 (i.e., 45 dots as "60", 60 dots as "80", and 75 dots as "100") was optimal in eliciting significant calibration effects in most undergraduate participants without raising strong suspicions about the feedback's accuracy. We adapted the feedback protocol used by Opfer and Siegler (2007) and told participants, "After you estimate the number of dots in each group, we will tell you how many dots there were, so you can see how close you were." During the feedback presentation, the dot array was not shown concurrently, so participants were made to calibrate their perceptual representations of numerosity based on the discrepancy between their estimates and the feedback. It is also crucial to note that we deliberately chose numerosities outside of the range in the test phase, so that any observed discontinuities in calibration effects within the tested range could not be attributed to the introduction of new associative mappings through the feedback given. The test phase comprising 310 trials was identical in every aspect to the uncalibrated estimation task, except that new arrangements of the dot sets were used to prevent the effects of configurational familiarity. This task was also self-paced and took about 20–30 min.

Questionnaire

Following the estimation tasks, participants completed an online questionnaire administered and managed using REDCap electronic data capture tools (Harris et al., 2009). Participants were asked whether they noticed anything unusual

about the estimation tasks, were probed about the strategies which they used, when they used each strategy, and the confidence in their estimates (see Appendix for the list of questions). To avoid the calibration manipulation influencing the strategy reports (or vice versa), delaying the strategy reports to the end of estimation tasks was necessary. As participants were only asked to report their strategies after both estimation conditions rather than immediately after each condition or on a trial-by-trial basis, it was likely that they would under-report, such as providing only the most salient strategy that came to mind. Hence, we opted for a semi close-ended report by cueing them with commonly used strategies. To this end, participants were allowed to select from multiple known strategies that were reported in previous studies (Gandini et al., 2010; Gandini, Lemaire, & Dufau, 2008). The four strategies included were: (a) Exact counting (e.g., I counted exactly in groups of *X*, and I am certain that there are *X* dots in each group.), (b) Approximate counting (e.g., I first saw one group of about *X* dots, a group of about *Y* dots, and another group of about Z dots. Therefore, there were about $(X+Y+Z)$ dots.), (c) "Benchmarking"—retrieved a quantity from memory, and approximately added or subtracted from it (e.g., I quickly looked at all the dots, thought it looked like there are about *X* or a little bit more, so I said slightly more than *X*.), (d) "Anchoring" (or "Decomposition/Recomposition", differences between them are subtle[1] and were considered as a single strategy here; see Gandini et al. 2010; Gandini, Lemaire, & Dufau, 2008)—Used a subset of the dots as an anchor (e.g., I counted exactly a group of *X* dots, or saw a group about *X* dots. Then, I estimated that there were six other similar groups, so I figured that there are 7*X* dots.). We also included (e) instinctively knew each and every quantity from memory, (f) no particular strategy that I was aware of, and (g) others (with open-ended responses solicited), to capture other possibilities. Note that numerosity-numeral mappings are referred to simply as "quantity/quantities" in the strategy descriptions that participants read as they are more intuitively understood.

[1] In Gandini, Lemaire, and Dufau's (2008) study, "anchoring" was defined as "Participants enumerated several dots (via counting), visually estimated the remaining dots based on the first enumeration, and then added the enumerated result and the estimated result" (e.g., "I first counted 3 dots, then 4 dots, added 3 and 4 = 7. Then, I estimated that there remained approximately twice as many dots, so I figured that there are 7 + 14 = 21 dots"). "Decomposition/recomposition" was defined as "Participants spotted one group of few dots, up to about four or five items, estimated the number of analogous groups, and then multiplied the number of items primarily subtilized [*sic*] by the estimated number of groups." (e.g., "I saw a group of 3 dots, and I estimated that there were six other similar groups; so I multiplied 7 by 3, and thought there are approximately 21 dots"). In the current study, we focus on the fact that subgroups of dots were used to enumerate the whole collection, regardless of whether participants used counting or subitizing to enumerate the subgroups, or whether they used multiplication or addition strategies.

Mathematical competence

Mathematical competence was measured using the Math Fluency and Calculation subtests of the Woodcock-Johnson III Tests of Achievement (WCJ-III; Woodcock, McGrew, & Mather, 2001). The Math Fluency subtest requires participants to solve 160 simple addition, subtraction, and multiplication problems with the numerals 0–10 as quickly as possible within 3 min (reliability = 0.92; Woodcock et al., 2001). Hence, this subtest primarily assesses fluency of arithmetic fact retrieval. The Calculation subtest is an untimed test including 45 items assessing arithmetic (with natural and rational numbers), algebra, trigonometry, and calculus (reliability = 0.89; Woodcock et al., 2001). Hence, it assesses a broader scope of calculation competence comprising procedural and conceptual knowledge. As there is growing evidence that the relation between numerosity-numeral mappings and math competence depends on the type of math competence assessed (Jang & Cho, 2018; Libertus et al., 2016; Yeo et al., 2019), the subtests scores were examined separately. Table 1 shows that the sample has a wide and representative range of math achievement scores. Age-normed standard scores were used for all analyses.

Control measures

Domain-general factors including the fluency of retrieving symbol–referent associations and working memory were also measured to assess the specificity of the relation between associative mapping mechanisms and mathematical competence. Participants' competencies in employing estimation strategies that rely on working memory (Gandini, Lemaire, Anton, & Nazarian, 2008) may be related to the strength and extent of associative mappings.

Reading competence The ability to infer non-numerical symbol–referent associations fluently was measured using the Reading Fluency subtest of the WCJ-III Tests of Achievement (reliability = 0.90; Woodcock et al., 2001). It requires participants to read a series of sentences and assess their truthfulness as quickly as possible within three minutes. Age-normed standard scores were used for all analyses.

Working memory Working memory capacity was measured using the composite performance on three complex span tasks: operation span, symmetry span, and rotation span (Foster et al., 2015). A composite score from two or more tasks that captures the shared variance among them is preferred to a single task score that captures variance that may be unrelated to working memory capacity (e.g., task-specific variance) (Foster et al., 2015). A single block of each task was administered in the order as mentioned. Compared to the typical use of three blocks per task, Foster

et al. (2015) found that a single block per task shortened the administration time by about 20 min, while not substantially reducing its predictive value of fluid intelligence (three blocks per task accounted for 51.1% of the variance in fluid intelligence, whereas one block per task still accounted for 46.5%). The reliabilities of the operation, symmetry, and rotation span tasks (1 block) are Cronbach's $\alpha = 0.69$, 0.61, and 0.66, respectively (Foster et al., 2015).

In each trial of the operation span task, participants had to remember a series of single letters and solve arithmetic problems (as distractors) that alternated with the to-be-memorized letters. There were 3–7 letters to be memorized in each trial and participants had to recall the letters in the order which they were presented. The total number of letters recalled in their correct order over the whole task (known as the partial score) was used as a measure of working memory capacity (Conway et al., 2005). There was a total of 25 letters to be remembered in a single block of the operation span task.

In each trial of the symmetry span task, participants had to remember a series of locations of single red squares in a 4-by-4 grid and judge the symmetry of shapes that alternated with the to-be-memorized locations of the squares. There were two to five locations to be memorized in each trial and participants had to recall the locations in the order which they were presented. There was a total of 14 locations to be remembered in a single block of the symmetry span task that contributed to the partial score.

In each trial of the rotation span task, participants had to remember both the length (short or long) and direction (one of eight cardinal and intercardinal directions) of a series of single arrows, and judge whether a rotated letter was laterally flipped (i.e., mirror image) or not that alternated with the to-be-memorized arrows. There were 2–5 locations to be memorized in each trial and participants had to recall the arrows in the order which they were presented. There was a total of 14 arrows to be remembered in a single block of the rotation span task that contributed to the partial score.

The partial scores for each task were transformed into z scores and then averaged to form a composite working memory score (e.g., Gonthier, Thomassin, & Roulin, 2016).

Analyses

Data management

Criteria for data exclusion of the test trials of the estimation task were adapted from previous studies that employed calibration paradigms (Alvarez et al., 2017; Sullivan & Barner, 2013, 2014; Yeo et al., 2019). Across 17,670 trials pooled from all participants, we first excluded null responses (uncalibrated: $N = 52$ trials; calibrated: $N = 40$ trials), and responses of "0" and "1" (uncalibrated: $N = 28$ trials; calibrated: $N = 63$ trials). Next, we excluded responses that were likely to be typing errors that are independent of idiosyncratic estimation abilities, specifically more than or equal to ten times as large ($10x$), and less than or equal to ten times as small as the numerosity (x) presented ($x/10$) (uncalibrated: $N = 33$ trials; calibrated: $N = 51$ trials). Previous research suggest that typical adults deviate from the true numerosity by up to a factor of at most four (e.g., Minturn & Reese, 1951; Yeo et al., 2019). Within each condition, we also excluded outlying estimates that were more than three times the median absolute deviation (MAD) of each participant's estimates of each numerosity presented (uncalibrated: $N = 891$ trials; calibrated: $N = 1021$ trials). MAD was computed from the median of the absolute deviations from the median and scaled by a constant $b = 1.4826$ assuming an underlying normal distribution (Leys, Ley, Klein, Bernard, & Licata, 2013). It is a robust measure of dispersion and is preferred to the use of standard deviations from the mean, which are themselves highly sensitive to outliers and hence less effective in detecting outliers (Leys et al., 2013). However, when more than 50% of the data points are identical (e.g., [5, 5, 5, 5, 5, 6, 6, 6, 6, 5] for numerosity 5), MAD will be equal to 0 and any data point that is not equal to the median (e.g., 6 in the example) will be flagged as an outlier, leading to false positives. In such cases, we flagged for outliers with an alternative measure to MAD, which uses the mean of the absolute deviations from the median, scaled by constant $b = 1.253314$ (see IBM's "modified z score"). Across the whole sample, 87.42–97.74% ($M = 94.32\%$) and 84.84–97.10% ($M = 93.35\%$) of each individual's uncalibrated and calibrated data points, respectively, were retained for further analyses (see Fig. S1 for scatterplots of all 57 participants). This resulted in a mean of 9.43 trials per numerosity (Mdn = 10, range = 5–10) for the uncalibrated condition, and a mean of 9.34 trials per numerosity (Mdn = 10, range = 5–10) for the calibrated condition.

Individual task-level effects of calibration

We first characterized the extent to which the calibration feedback was effective in inducing changes in participants' estimates. Following previous studies (Alvarez et al., 2017; Sullivan & Barner, 2013, 2014; Yeo et al., 2019), we determined whether each participant was influenced by calibration by regressing their estimates on numerosity (5–35; as a continuous variable), calibration condition (uncalibrated, calibrated), and the numerosity by calibration interaction using $p < 0.05$ based on the F test related to each effect. A participant was classified as a "calibrator" when there was either a significant main effect of calibration or an interaction between calibration and numerosity (Alvarez et al., 2017; Sullivan & Barner, 2013, 2014; Yeo et al., 2019). We assumed unequal variances across numerosities and

conditions, used Type II sums of squares (Langsrud, 2003) and heteroscedasticity-robust standard errors to make statistical inferences [using R packages *lmtest* (Zeileis & Hothorn, 2002) and *Car* (Fox & Weisberg, 2011)].

Estimation indices

We computed participants' accuracy, consistency, and linearity of the uncalibrated and calibrated estimates separately, and estimated their base range of associative mappings (Table 1). We note that the computation of the accuracy, consistency, and linearity indices below are non-independent ($r_s = 0.58–0.79$, Table 2). We considered all of them in the current study only because previous studies have used them to establish relations between estimation performance and math competence, and one of the aims of this study was to examine whether the base range of associative mappings could underlie these relations. Moreover, we are interested in estimation performance in general, rather than the unique relations associated with each performance index. We assessed their split-half reliabilities by computing each index from the first and second halves of the trials for each numerosity across all numerosities, and then performed a Spearman–Brown Prophecy correction to predict its full-length reliability.

Accuracy of estimates Accuracy was indexed by averaging the absolute error rates across all data points ($\text{AER} = \frac{|\text{Estimate - Numerosity}|}{\text{Numerosity}}$) (e.g., Alvarez et al., 2017). The absolute value avoids potential reciprocal cancelation between under- and over-estimation during averaging, and retains information about trial-level accuracy. A smaller AER thus reflects greater accuracy at the task level. Split-half reliability was high in both the uncalibrated condition [Spearman's rho, $r_s(55) = 0.80$, $p < 0.001$, Spearman–Brown = 0.89] and calibrated condition [$r_s(55) = 0.93$, $p < 0.001$, Spearman–Brown = 0.96].

Consistency of estimates Consistency was indexed by computing the coefficient of variation per numerosity ($\text{CV} = \frac{\text{Standard deviation of estimates}}{\text{Mean estimate}}$) and taking the mean CV across the range of target numerosities. A smaller CV reflects more consistent estimates at the task level. Split-half reliability was high in both the uncalibrated condition [$r_s(55) = 0.77$, $p < 0.001$, Spearman–Brown = 0.87] and calibrated condition [$r_s(55) = 0.90$, $p < 0.001$, Spearman–Brown = 0.94].

Linearity of estimates The extent to which participants' estimates were proportionally spaced relative to one another was indexed by R^2_{lin} of a simple linear regression model with participant's trial-level estimates regressed on numerosity as a continuous variable (e.g., Alvarez et al., 2017; Sullivan, Frank, & Barner, 2016). A higher R^2_{lin} reflects more proportionally spaced estimates. Split-half reliability was high in both the uncalibrated condition [$r_s(55) = 0.70$, $p < 0.001$, Spearman–Brown = 0.82] and calibrated condition [$r_s(55) = 0.74$, $p < 0.001$, Spearman–Brown = 0.86].

Individual base range of associative mappings Using an approach employed by previous studies (Alvarez et al., 2017; Sullivan & Barner, 2013, 2014; Yeo et al., 2019), sequential pairwise comparisons per numerosity were performed and the smallest numerosity that was influenced by calibration was defined as the onset of structural mapping. Correspondingly, we defined the upper bound of the base range of associative mappings in each individual as one integer less than the smallest numerosity that was influenced by calibration (i.e., if numerosity 7 was the smallest numerosity that showed an effect of calibration using this sequential approach, numerosity 6 was the upper bound of the base range of associative mappings). To measure the base range of associative mappings, for each participant, we performed a series of Welch's *t* tests on the estimates in the uncalibrated and calibrated conditions for each numerosity using $p < 0.05$ as the statistical significance threshold. Welch's *t* test does not require the assumption of equal variances between conditions to hold, unlike Student's *t* test (Delacre, Lakens, & Leys, 2017). As we were interested in the continuity of any statistically significant calibration effects, we did not correct for multiple comparisons, which may introduce discontinuities when some small but true calibration effects are considered false positives. Using this approach, two participants showed an effect of calibration even for numerosity 5. As we assumed that participants should be able to subitize sets of four or less and that these sets are likely to be unaffected by calibration, we imputed "4" as the upper bound for these two participants. Results were qualitatively similar even if we excluded these two participants from analyses that involved the base range measure, rather than imputing the base range. Split-half reliability analysis with the base range of associative mappings computed from the first and second halves of the data was close to acceptable, $r_s(53) = 0.53$, $p < 0.001$, Spearman–Brown = 0.69.

To account for any potentially non-normal distributions resulting from the small number of data points in some cases, we also replicated the analyses using a non-parametric version of the Welch's *t* test on ranked data (Zimmerman & Zumbo, 1993). As the estimated base range of associative mappings using both methods were almost perfectly correlated [$r_s(55) = 0.99$, $p < 0.001$] and the results were qualitatively similar, we only report the results using the raw data. This suggests high consistency in the base range of associative mappings regardless of parametric and non-parametric approaches.

Univariate normality

Descriptive statistics of standardized math and reading measures, working memory measures, and estimation indices are presented in Table 1. Most key indices were normally distributed (see Table 1; Shapiro–Wilk; all $ps > 0.052$). For non-normal variables, we performed a rank-based inverse normal transformation (Bishara & Hittner, 2012, 2015) on them before performing any correlational analyses. All non-normal variables became normally distributed upon transformation (Shapiro–Wilk; $ps > 0.11$, |skewness| < 0.14, |kurtosis| < 0.23). Regardless, the absolute skewness and kurtosis of all variables are within the acceptable ranges for standard parametric analyses (< 2 for skewness and < 7 for kurtosis) (Byrne, 2010; Hair Jr., Black, Babin, & Anderson, 2010; Kline, 2011).

Correlation analyses

To address whether individual differences in the base range of associative mappings were related to estimation performance, as well as with math competence, we conducted zero-order and partial correlational analyses. Relations with performance indices for the calibrated estimates were not analyzed (other than with their uncalibrated counterparts), because they may be dependent on calibration value and method (i.e., they may not generalize to other studies), and are thus not of theoretical interest here. For each set of related tests addressing a specific question (i.e., appearing in separate sections of the results), we used Benjamini and Hochberg's (1995) false discovery rate procedure to correct for multiple comparisons. Uncorrected *p* values are reported, and unless otherwise noted, all significant correlations reported remained significant after correction.

Analyses of questionnaire items

Our primary analyses focused on the quantitative analyses of frequencies of strategy choice using Chi-square and binomial tests of proportions. For exploratory qualitative analyses, participants' responses were coded by the first author and the complete set of raw qualitative responses are provided in the Online Resource.

Bayesian analyses

To provide measurable evidence in support of positive, inconclusive, and null findings within this dataset (Dienes, 2014), we conducted complementary Bayesian *t* tests, correlational analyses, and frequency analyses using JASP 0.9.0.1 (JASP Team, 2019), jamovi 0.9.2.3 (The jamovi project, 2019) and their default "objective" priors (Cauchy distribution scaling factor $r = 0.707$ for *t* tests, stretched beta prior width = 1 for correlation tests, and beta priors = 1 for tests of frequencies/proportions). Whenever possible, we report the Bayes factor (BF_{10}), which indicates the likelihood that the evidence is in favor of the alternative hypothesis relative to the null hypothesis (Wagenmakers, Love, et al., 2018; Wagenmakers, Marsman, et al., 2017). For instance, a BF_{10} of 3 suggests that the data were three times more likely to occur under the alternative than the null hypothesis. BFs greater than 3, 10, 30, and 100 are considered "moderate", "strong", "very strong", and "extreme" evidence in support of the alternative hypothesis (Jeffreys, 1961; Lee

Table 1 Descriptive statistics of pre-transformed standardized achievement, working memory, and estimation measures ($N = 57$)

Measure	Mean	Median	SD	Range	Skewness	Kurtosis
WCJ-III Calculation	124.23	126	13.27	92 to 147	−0.47	−0.24
WCJ-III Math Fluency	114.21	115	12.18	91 to 146	0.20	−0.11
WCJ-III Reading Fluency	118.72	120	12.37	92 to 147	−0.03	−0.19
Operation span	21.47	23	3.60	8 to 25	−1.41	2.54
Symmetry span	10.54	11	2.66	4 to 14	−0.76	−0.10
Rotation span	9.68	9	3.00	4 to 14	−0.20	−0.93
Composite working memory	0[a]	0.12	0.73	−1.95 to 1.24	−0.54	−0.34
Uncalibrated AER	0.18	0.17	0.05	0.08 to 0.31	0.74	0.52
Calibrated AER[b]	0.30	0.20	0.22	0.07 to 1.24	2.12	5.21
Uncalibrated CV[b]	0.14	0.14	0.04	0.07 to 0.25	0.88	1.58
Calibrated CV[b]	0.18	0.17	0.06	0.08 to 0.38	1.28	2.22
Uncalibrated R^2_{lin}	0.76	0.76	0.08	0.53 to 0.91	−0.47	0.06
Calibrated R^2_{lin} [b]	0.73	0.74	0.10	0.35 to 0.90	−1.34	3.20
Base range of associative mappings[b]	9	8	4.53	4 to 22	1.27	0.87

WCJ-III Woodcock–Johnson III Tests of Achievement, *AER* absolute error rate, *CV* coefficient of variation

[a]Composite working memory scores are *z* scores

[b]These measures are not-normally distributed (Shapiro–Wilk; all $ps < .006$)

& Wagenmakers, 2013; Wagenmakers et al., 2018). We adopted Dienes' (2014) criteria of BFs greater than 3 or less than 1/3 as conclusive evidence in support of the alternative or null hypothesis, respectively, and BFs between 1/3 and 3 (inclusive) as inconclusive evidence.

Results

Task-level effects of numerosity and calibration, and manipulation success

The calibration manipulation was meant to induce an increase in a linear slope of the estimates by 0.33. The mean within-participant change in slope was + 0.39 (Mdn = + 0.28). All 57 participants showed a main effect of numerosity, 55 (96.5%) showed a main effect of calibration, 50 (87.7%) showed a calibration by numerosity interaction, 56 (98.3%) showed either a main effect of calibration or a calibration by numerosity interaction [hereafter referred to as "calibrators" (e.g., Alvarez et al., 2017; Sullivan & Barner, 2014)], and 49 (86.0%) showed both main effect of calibration and a calibration by numerosity interaction (see Fig. S1 for individual plots). All proportions reported above were above chance (0.05 for each independent effect, 0.1 for either main effect or interaction, and 0.025 for both main effect and interaction) [one-sided binomial $ps < 0.001$, all BFs $> 3.73 \times 10^{52}$].

Twenty-three participants (40.4%) reported that they noticed something unusual about the estimation tasks. A primary concern regarding manipulation failure was whether this subset of participants ignored the manipulation and showed no calibration effects. However, their noticing was largely independent of whether they were influenced by the manipulation as all but one participant was affected by the calibration ($\chi^2 = 0.69$, $df = 1$, $\phi = 0.110$, $p = 0.407$, $BF_{10\ Poisson} = 0.22$; see Online Resource Table S1). Moreover, of those 23 participants, most participants merely expressed uncertainty about the calibration feedback relative to the perceived numerosity rather than the objective numerosity that was unbeknownst to them (e.g., "when told the correct number of dots, they seemed much higher than the amount of dots I saw") and only four participants expressed with some certainty suspicions about the accuracy of the calibration feedback (i.e., "I believe the 'correct' numbers that I was given were wrong. They were way too high") (see Online Resource). Taken together, there is no strong evidence that participants' knowledge about the objective numerosity during the calibration phase affected how they made use of the feedback for the post-calibration phase.

Individual base range of associative mappings

There was considerable inter-individual variability in how high the base ranges of associative mappings were (range = 4–22) (Fig. 2 and Table 1). As shown in Fig. 2, the distribution was positively skewed with numerosity 5 as the modal upper bound.

Relation between uncalibrated estimation performance and math competence

Participants with higher Calculation scores had more accurate [$r(55) = -0.40$, $p = 0.002$, $BF_{10} = 16.98$] and more linear [$r(55) = 0.43$, $p < 0.001$, $BF_{10} = 37.26$] estimates. They also tended to have more consistent estimates [$r(55) = -0.32$, $p = 0.015$, $BF_{10} = 2.96$], although the evidence is weaker compared to accuracy and linearity. There is, however, weak-to-moderate evidence that accuracy, consistency, and linearity of estimates were not related to Math Fluency [all $ps > 0.13$, all BFs < 0.49, Table 2].

Most of these findings held after controlling for Reading Fluency and working memory. An exception is that consistency of estimates was correlated with Math Fluency [$r(53) = -0.29$, $p = 0.030$, $BF_{10} = 1.66$], even though it was not before controlling for these confound variables. Nonetheless, Bayesian analysis suggests that the evidence is weak.

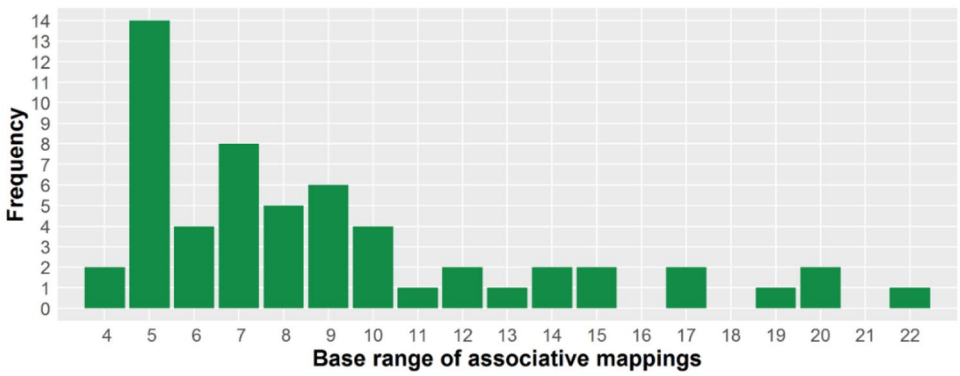

Fig. 2 Distribution of participants' base range of associative mappings (see Table 1)

Relations involving base range of associative mappings

Individual differences in the base range of associative mappings were not related to estimation performance, calculation skills, and arithmetic fact retrieval fluency [all $ps > 0.36$, all BFs < 0.25, Table 2]. These findings remained unchanged after controlling for Reading Fluency and working memory [all $ps > 0.16$, all BFs < 0.45, Table 2].

Finally, in the relation between estimation performance and math competence, we further controlled for the base range of associative mappings to probe whether it functioned as a confounding third variable. Calculation remained correlated with accuracy [$r(52) = -0.34$, $p = 0.012$, $BF_{10} = 3.55$], consistency [$r(52) = -0.32$, $p = 0.017$, $BF_{10} = 2.67$] and linearity [$r(52) = 0.40$, $p = 0.003$, $BF_{10} = 13.44$] of estimates. Math Fluency also remained correlated with estimation consistency [$r(52) = -0.30$, $p = 0.030$, $BF_{10} = 1.67$], although the evidence is still weak and inconclusive.

In sum, estimation performance was related to Calculation (but not Math Fluency) over and above Reading Fluency, working memory, and the base range of associative mappings. Importantly, there is evidence that the base range of associative mappings was not related to estimation performance or any of the math competence measures, and is, therefore, unlikely to underlie the relation between estimation performance and math competence.

Continuity of calibration effects

In previous studies, it was assumed that all numerosities beyond this upper bound would show calibration effects, and that these effects are a signature of the interdependency among structural mappings. In other words, it was assumed that once structural mapping was initiated, all subsequent numerosities would be structurally mapped and susceptible to calibration. However, when we extended the sequential tests beyond this upper bound to the largest numerosity presented (i.e., 35), calibration effects were notably discontinuous (see Figs. 3 and S2). Only three participants showed continuity in calibration as predicted by the response-grid model. In other words, discontinuity in calibration effects (54 out of 57 participants, 94.7%) was very much the norm rather than an exception (chance = 0.5, one-sided binomial $p < 0.001$, $BF_{10} = 1.70 \times 10^{11}$).

We further examined the trends for standardized calibration effect sizes (Cohen's *d*) in each participant. In Fig. S3 of the Online Resource, the distributions of effect sizes of the statistically significant and non-significant calibration effects did not seem to overlap substantially. Due to the huge imbalance in significant and non-significant effects within most participants, we computed instead the participant-specific difference in median absolute effect sizes of the statistically significant and non-significant effects, and then assessed the difference in effect size distributions at the group level. The

Table 2 Zero-order and partial correlation coefficients of estimation indices, standardized achievement test scores, and working memory ($N = 57$)

Measure	1	2	3	4	5	6	7	8
1. Base range of associative mappings	–	−0.103	−0.124	0.033	−0.119	−0.100	−0.168	0.258
BF_{10}		(0.22)	(0.25)	(0.17)	(0.24)	(0.22)	(0.35)	(1.04)
2. Uncalibrated AER	0.012	–	0.575***	−0.730***	−0.400**	−0.105	0.014	−0.431***
BF_{10}	(0.17)		(7181.37)	(1.02×10^8)	(16.98)	(0.22)	(0.17)	(38.33)
3. Uncalibrated CV	−0.054		–	−0.793***	−0.321*	−0.200	0.182	−0.168
BF_{10}	(0.18)			(4.22×10^{10})	(2.96)	(0.49)	(0.40)	(0.35)
4. Uncalibrated R^2_{lin}	−0.043				0.429***	0.0123	−0.082	0.228
BF_{10}	(0.18)				(37.26)	(0.25)	(0.20)	(0.68)
5. Calculation	−0.193	−0.334*	−0.306*	0.400**	–	0.318*	0.046	0.255
BF_{10}	(0.44)	(3.46)	(2.07)	(14.53)		(2.79)	(0.17)	(0.98)
6. Math Fluency	−0.037	−0.153	−0.292*	0.175		–	0.326*	−0.045
BF_{10}	(0.17)	(0.31)	(1.66)	(0.37)			(3.28)	(0.17)
7. Reading Fluency							–	0.007
BF_{10}								(0.17)
8. Working memory								–
BF_{10}								

Upper triangle consists of zero-order correlations. Lower triangle consists of partial correlations controlling for reading fluency and working memory

AER absolute error rate, *CV* coefficient of variation

*$p < 0.05$. **$p < 0.01$. ***$p < 0.001$. BF_{10} = Bayes factor (alternative/null hypotheses)

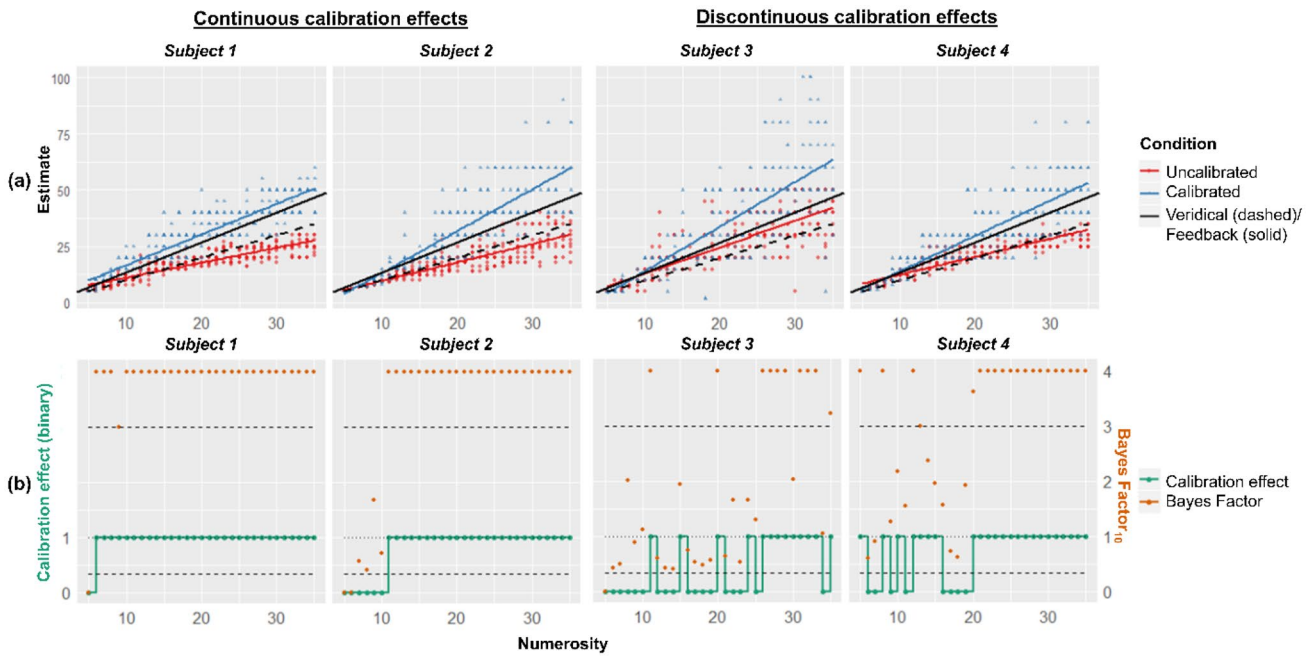

Fig. 3 Examples of continuous and discontinuous calibration effects in four participants who showed a calibration by numerosity interaction effect. **a** Raw estimates per condition. **b** Green step plots reflect binary coding of "1" for significant pairwise difference between conditions per numerosity and "0" for non-significant pairwise difference at $p < 0.05$ (uncorrected) for the corresponding plots above. Orange data points represent the Bayes factors artificially bounded between 0 and 4. Bayes factors > 3 and < 1/3 (dashed lines) reflect evidence in favor of a calibration effect and lack thereof, respectively. Bayes factors close to 1 (dotted line) reflect data insensitivity in distinguishing between the null and alternative hypotheses. The continuity of the calibration effects and trend of effect sizes for all 57 participants are shown in Figures S2 and S3

median within-participant difference in effect sizes between significant and non-significant effects was 0.91 ($M = 1.02$, range = 0.46–3.05), Wilcoxon signed-rank test, $W = 1653$, $p < 0.001$. These large-effect size differences suggest that the discontinuities observed were unlikely to be driven by statistical randomness.

We also assessed the evidence for strong associative mappings among the 822 non-significant effects out of 1767 effect sizes in the whole sample. Out of the 822 non-significant effects, 96 (11.7%) are true null differences (e.g. participants responded "5" consistently for numerosity 5 regardless of calibration). Most of these true null differences are observed for numerosities 10 or less (see Online Resource Table S3), which is consistent with the notion that calibration effects are dependent on set size hypothesized by Sullivan and Barner (2013, 2014). At the individual level, 10 participants had true null differences for 5 and 6; 9 participants had true null differences for 5 through 7; 2 participants had true null differences for 5 through 8; 1 participant each had true null differences from 5 through 9, and 5 through 10. Taken together, these true null differences provide support for the presence of strong associative mappings among the small numerosities that are unlikely to be based on the same mechanism as mappings for larger numerosities.

Even though the trend of the calibration effects across all numerosities should be the primary focus, there is still a concern that the observed discontinuities may be due to a lack of statistical power for individual numerosities, with at most 10 data points per calibration condition. We thus explored whether these discontinuities still exist with a greater number of data points per calibration condition using a "sliding window" of 2, 3, or 5 numerosities. This approach is theoretically valid given the approximate nature of the analog number system (Izard & Dehaene, 2008). For instance, a sliding window of 3 numerosities included 5, 6, and 7 in the first window, 6, 7, and 8 in the second window, 7, 8, and 9 in the third window, and so on. For the analyses with sliding windows of 2, 3, and 5 numerosities, 54, 53, and 54 out of 57 participants, respectively, showed discontinuities in the calibration effects (chance = 0.5, one-sided binomial ps < 0.001, BF_{10}s > 1.26×10^{10}). Even at these coarser resolutions with presumably higher statistical power (two- to

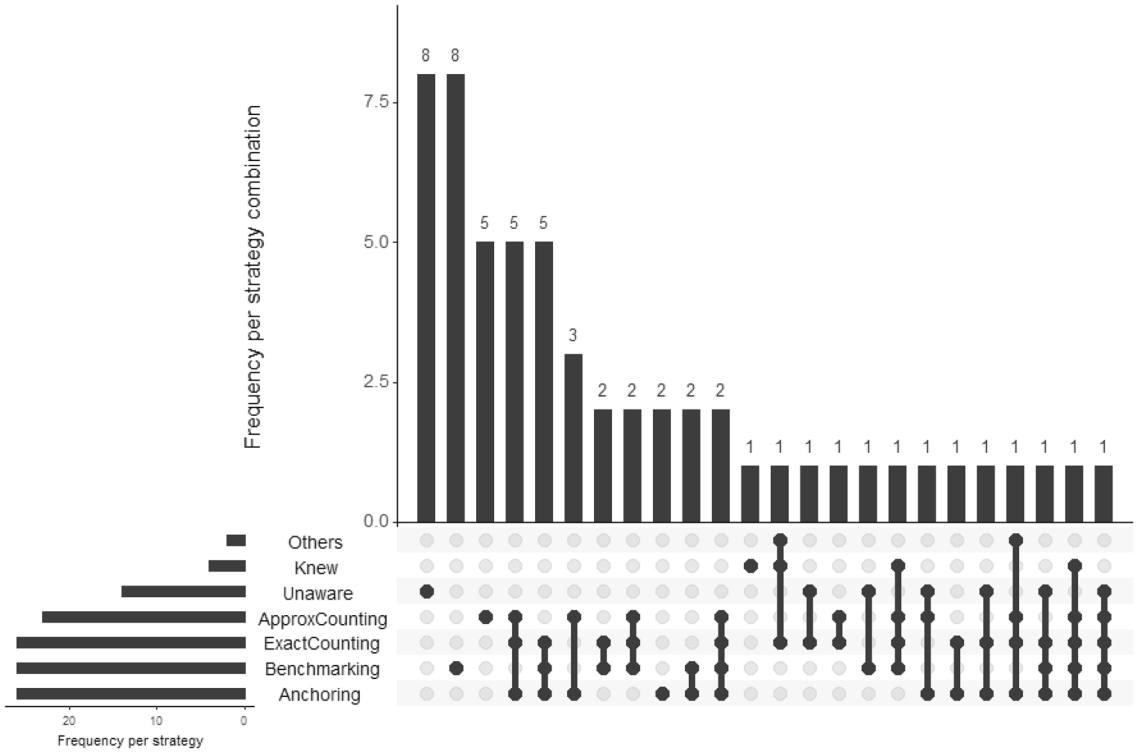

Fig. 4 Summary of the frequencies for combinations of strategy used during the uncalibrated estimation task. The choices that participants could choose from include: **a** using a subset of the dots as an anchor ("Anchoring"), **b** retrieved a quantity from memory, and approximately added or subtracted from it ("Benchmarking"), **c** exact counting ("ExactCounting"), **d** approximate counting ("ApproxCounting"), **e** instinctively knew each and every quantity from memory ("Knew"), **f** no awareness of particular strategy ("Unaware"), and **g** others

fivefold increase), the qualitative patterns of the results remained unchanged (see Figs. S5–S9).[2]

Strategy use during uncalibrated estimation

A hypothesis from the response-grid model is that in every individual, a substantial number of estimates or numerosity-numeral mappings will be retrieved from a response grid. In other words, it predicts that most if not all participants will use benchmarking as part of their strategy repertoire at least on some of the trials. Hence, we focused most of our analyses on the use of benchmarking, but provided a detailed summary of the questionnaire data and other exploratory analyses in the Online Resource. We first excluded participants who were metacognitively unaware of how they performed the task and did not choose any other strategy options ($N = 8$). Hence, the following analyses were conducted for the remaining 49 participants. The proportion of participants who used more than one strategy (33/49, 67%, median number of strategies = 3) was greater than chance (0.5, binomial $p = 0.01$, $BF_{10} = 6.67$) (Fig. 4). By and large, participants varied the use of strategies in their repertoire based on the set size and the configuration of dot sets (see Online Resource). It is clear from Fig. 3 that the proportion of participants who used benchmarking as part of their strategy repertoire (26/49, 53%) was not greater than chance (0.5, binomial $p = 0.39$, $BF_{10} = 0.26$). Even if we included participants who reported that they instinctively knew each and every numeral-numerosity mapping from memory (but not also benchmarking) (2/49, 4.1%), the proportion (28/49, 57%) was still not greater than chance (0.5, binomial $p = 0.20$, $BF_{10} = 0.48$). No other strategy was greater than chance either (0.5, binomial $ps > 0.38$, $BFs < 0.26$). The proportions of participants using the four strategies most commonly used: benchmarking (26/49, 53%), anchoring (26/49, 53%), exact counting (26/49, 53%), and approximate counting (23/49, 50%) were not significantly different from one another ($\chi^2 = 0.55$, $df = 3$, $p = 0.907$), suggesting that there was no single strategy that was essential in any participant's strategy repertoire. Of the 26 participants who used benchmarking, only six (23%)

[2] We also used the base range of associative mappings estimated from the "sliding window" analysis of 2 numerosities to assess the consistency of the original base range index. They were highly correlated [$r_s(54) = 0.74$, $p < 0.001$], indicating that individual differences captured by the original index are consistency across estimation approaches.

mentioned that there were particular numeral-numerosity mappings that they retrieved from memory. In other words, when benchmarking was used, it was more likely that there were no particular benchmarking numeral-numerosity mappings retrieved (chance = 0.5, binomial $p = 0.005$, $BF_{01} = 21.53$), possibly due to the use of numeral-numerosity mappings from previous trials as benchmarks (e.g., "guessed off of memory of previous trials", see Online Resource for explicit mention by at least three participants).

Finally, we focused on the three participants who showed continuity in their calibration effects (see Fig. S2) and examined their reported strategy use. One participant was not particularly aware of the strategies used, one used only anchoring (anchor subsets: 5, 10), and one used exacting counting, benchmarking (benchmark numerosity-numeral mappings: 10, 15, 20, 25), and anchoring (anchor subsets: 5, 10). Taken together, there is little evidence that is consistent with the hypothesis that a response grid is even used at all by most of our participants.

Exploratory analyses relating strategy use and degree of consistency in estimation

We explored the extent to which estimation performance was constrained between the calibrated and uncalibrated conditions. If a response grid is primarily used for estimation in both conditions, the degree of consistency of estimates should be highly constrained by the supposedly fixed acuity of the analog mental number line to which the response grid is mapped onto. If multiple strategies are typically used, and may differ between conditions, the degree of consistency should be less constrained between conditions. In other words, any change in consistency is likely due to external factors such as strategies than the acuity of representations per se. On average, CVs were higher in the calibrated condition ($Mdn = 0.17$) than in the uncalibrated condition ($Mdn = 0.14$) (Wilcoxon signed-rank test, $W = 1457$, $p < 0.001$). There was no conclusive evidence that participants who reported that their strategies differed after calibration (27 out of 57, 47.4%) had a greater change in CV ($M = 0.041$, $Mdn = 0.034$, $SD = 0.041$) than those who reported that their strategies did not differ ($M = 0.032$, $Mdn = 0.012$, $SD = 0.053$) (Mann–Whitney U test, $W = 322$, $p = 0.189$, $BF_{10} = 0.44$). An individual differences approach may also provide insights into the idiosyncrasies in changes in consistency (i.e., reduced in some, but increased in others), possibly due to strategy-related changes. Previous studies showed that when drastic strategic differences between conditions were not expected,[3] CVs (or equivalently, the Weber's fraction) were highly correlated between calibrated and uncalibrated conditions ($r = 0.82$, Yeo et al., 2019), or between two calibrated conditions ($r = 0.78$, Izard & Dehaene, 2008). In the current study, the explicit calibration was more likely to induce changes in strategy deployment, and indeed, the CVs were less strongly correlated between conditions [$r(55) = 0.58$, $p < 0.001$, $BF_{10} = 1.04 \times 10^4$] (see Table S2 for the full correlation matrix). Fisher's z comparison of the correlation coefficients in Yeo et al. (2019) and in the current study confirmed the attenuation, $z = 2.60$, $p = 0.0046$.

Discussion

Previous research suggests that the quality of numeral-numerosity mappings, typically indexed by performance during an estimation task in terms of accuracy, consistency and linearity of estimates, is related to math competence in children and adults (Alvarez et al., 2017; Bartelet et al., 2014; Booth & Siegler, 2006; Castronovo & Göbel, 2012; Chesney et al., 2015; Guillaume et al., 2016; Libertus et al., 2016; Lyons et al., 2014; Mazzocco et al., 2011a; Mejias, Grégoire, et al., 2012; Mejias, Mussolin, et al., 2012; Mejias & Schiltz, 2013; Pinheiro-Chagas et al., 2014; Wong et al., 2016a). The current study aimed to examine the mechanisms that underlie such a relation by focusing on the relative extents of two hypothesized mechanisms of mapping—item-based associative mapping and holistic structural mapping. Using fine-grained sampling during an estimation task, we were able to assess individual differences in the base range of associative mappings in adults, and examine whether that base range of associative mappings was related to estimation performance as well as to math competence.

Robust relation between estimation performance and math competence in adults

Our findings indicate that estimation performance is related to math competence even with a fine-grained sampling (31 numerosities between 5 and 35), and as such, are consistent with previous studies relating estimation performance to math competence in adults, all of which used a sparse sampling of numerosities (Castronovo & Göbel, 2012; Chesney et al., 2015; Guillaume et al., 2016; Mejias, Grégoire, et al., 2012). This suggests that the estimation–math relation is rather robust to the tested range and sampling procedure. Although previous studies typically controlled for reading ability and intelligence, strategy-related factors such as working memory are rarely controlled, at least in adult studies. Consistent with findings that several commonly employed estimation strategies require working memory resources (e.g., using subsets as anchors; Gandini

[3] Calibration method was either implicit (merely providing an upper bound in Yeo et al., 2019), or identical between two calibrated conditions (merely changing the calibration reference value in Izard & Dehaene, 2008).

et al., 2010; Gandini, Lemaire, Anton, et al., 2008; Gandini, Lemaire, & Dufau, 2008), we found that higher estimation accuracy was indeed related to higher working memory capacity. Nonetheless, the estimation–math relation remained after controlling for reading fluency and working memory, suggesting that the fluency of retrieving domain-general symbol–referent associations and working memory do not fully account for the estimation–math relation. Similarly, Wong et al. (2016a) found that linearity of estimates still provided unique contribution to children's arithmetic performance after controlling for visuo-spatial working memory and central executive function.

The estimation–math relation, however, seems to be relevant for procedural and conceptual calculation skills, but not fluency in the retrieval of overlearned arithmetic facts. This adds to the growing list of studies that have found differential relations between quality of numeral-numerosity mappings and distinct math sub-domains or skills (Brankaer et al., 2014; Holloway & Ansari, 2009; Jang & Cho, 2016, 2018; Libertus, Feigenson, & Halberda, 2013; Libertus et al., 2016; Lourenco, Bonny, Fernandez, & Rao, 2012; Mazzocco et al., 2011a; Mejias & Schiltz, 2013; Orrantia et al., 2019). For instance, Libertus et al. (2016) found that, in children, estimation consistency was related to formal math skills (e.g., arithmetic facts, place value), but not informal math skills (e.g., counting with fingers). Hence, it is critical for future studies to consider math competence as a composition of distinct constructs and skills rather than a holistic construct. In sum, the estimation–math relation is robust to sampling procedure, independent of domain-general processes such as working memory, and specific to mathematical concepts and procedures beyond arithmetic fact retrieval fluency.

Base range of associative mappings among small numbers is neither related to estimation performance nor to math competence

By comparing participants' estimates before and after calibration using a sequential testing approach for every numerosity tested, the current study is the first to precisely measure the base range of associative mappings in our participants. We found conclusive evidence that the base range of associative mappings neither relates to any of the estimation performance indices nor to any math measures. The base range also does not seem to account for the relation between estimation performance and procedural calculation skills. The lack of a relation is contrary to our prediction, but supports Sullivan and Barner's (2014) hypothesis that associative mappings play a negligible role in both estimation and math skills.

Discontinuity in calibration effects is inconsistent with a holistic structural mapping

It is thought that making numerical estimates involves a system-level mapping between the analog mental number line and a symbolic response grid (Izard & Dehaene, 2008). It is further thought such a holistic mapping between the two systems underlies most of the numerosity-numeral mappings, and strong associative mappings are limited only to small numbers (for a review, see Carey & Barner, 2019; Sullivan & Barner, 2013, 2014). Based on these key assumptions, although we could identify a base range of associative mappings in every participant, we found discontinuities in the calibration effects in majority of our participants. This replicates and extends our previous finding (Yeo et al., 2019) with a vastly different sampling range (5–35 vs. 8–350) and resolution (all vs. selected numbers within the range), and calibration paradigm (explicit vs. implicit). Multiple approaches used to analyze the trends in the calibration effects suggest that the discontinuities cannot be entirely explained by randomness due to a lack of statistical power. Moreover, the calibration method used here is highly similar to that used by Izard and Dehaene (2008), but with several modifications that ought to favor an observation of continuous calibration effects: (1) instead of one calibration inducer presented just once in Izard and Dehaene (2008), we used three calibration inducers with a fixed feedback-to-actual ratio presented three times each to instill and strengthen a more linear- or ratio-based calibration; (2) instead of providing a calibration inducer within the tested range in Izard and Dehaene (2008), we provided calibration inducers that are outside the tested range to avoid the possibility that memory of the calibration inducers would introduce inconsistencies in the effect sizes across numerosities, which could result in artefactual discontinuities. Despite these modifications, discontinuities in calibration effects are still observed. These discontinuities may stem from several non-mutually exclusive possibilities that can inform the existing theories of numerosity estimation and the mappings between numerical symbols and their magnitude referents more generally.

First, some strong associative mappings may exist amidst structural mappings of larger numbers. There is evidence consistent with this possibility. In previous studies that employed sparse sampling (8, 12, 20, … 350) with adult participants, the highest base range extended to 150 (Sullivan & Barner, 2013) or even 200 (Yeo et al., 2019). However, it is possible that the previous studies might have captured sparse associative mappings among large numbers rather than a truly continuous range of associative mappings. For instance, an individual who showed no calibration effects for 8, 12, and 20 previously has been assumed to have a

continuous set of associative mappings up to 20, but he or she might have shown some calibration effects from 13 through 19 if those had been sampled. In the current study, we sampled continuously up to 35 and the highest base range of associative mappings in any participant extended to only 22. If there are indeed strong associative mappings for larger numbers sparsely distributed across the mental number line (Dehaene & Mehler, 1992), then it may be that the whole collection of associative mappings for both small and large numbers, and not just the "base range" that we have delineated here, may play a critical role in estimation performance or math competence. Independent of the theories discussed, however, the base range of associative mappings in itself may have other theoretical value that are not explored within the scope of the current study, so we do not consider this novel measure to be entirely meaningless in light of the discontinuity finding. For instance, it can inform whether individual differences in the acuity of perceiving a difference of one between small sets (e.g., 6 vs. 7) are related to individual differences in the subitizing range and visuo-spatial working memory more generally (e.g., Piazza et al., 2011; Revkin et al., 2008).

Moreover, our calibration manipulation, which used values beyond the actual range assessed (e.g., 75 dots as "100" when the largest set tested had 35 dots), also provided evidence to support the idea that associative mappings are not exclusive to small numbers. Specifically, many participants made calibrated estimates that are close to what we would expect based on the calibration values (see Fig. S1), which suggests that not only can association-based mappings be formed for large numbers, but they can also be applied fairly consistently and accurately during an estimation task over the course of 20–30 min. There is also evidence that new associative mappings formed for up to sets of 210 can be so strong that they persisted even after eight months (Minturn & Reese, 1951). Hence, although strong associative numerosity-numeral mappings for large numbers may be rare in nature due to the lack of opportunities to learn those mappings, they are not difficult to acquire in lab-based novel symbol training studies (Lyons & Ansari, 2009; Lyons & Beilock, 2009; Malone, Heron-delaney, Burgoyne, & Hulme, 2019; Merkley & Scerif, 2015; Merkley, Shimi, & Scerif, 2016; Zhao et al., 2012). In any case, the possibility that associative mappings can exist for large numbers implies that calibration may not affect the full length of a response grid, but segments of it. In other words, a response grid may be composed of segments that are pieced together with different anchor points rather than a single anchor point, in which case, the transformation of any response grid may not be adequately characterized by a two-parameter (translation and scaling) model as proposed by Izard and Dehaene (2008).

Alternatively, it is possible that the discontinuities beyond the base range do not indicate the presence of strong associative mappings, but that system-level structural mappings may not exist, at least not in the way that the analog mental number line and a symbolic response grid are thought to be holistically linked. With a holistic mapping between the analog and symbolic systems (see Fig. 1), to the extent that a set of 25 dots has been calibrated to map to a different segment of the response grid, and that sets of 23 through 27 are not perceptually discriminable from sets of 25, there is no reason that calibration effects of comparable size should not also be observed for numerosities 23 through 27. Yet, the calibration effects do not show strong interdependence in most individuals. Such weak interdependency among adjacent numeral-numerosity mappings suggests that participants do not make use of a response grid during an estimation task, at least not primarily. If numerosities are processed instantaneously and in parallel as commonly assumed, and the response grid is the primary mechanism for making estimates, the retrieval-based benchmarking strategy should be in the repertoire of most if not all participants, even if it is not used all the time. However, only about half of our participants reported using benchmarking at all. It might be argued that our stimuli were presented for 500 ms, which was much longer than the 100 ms presentation used by Izard and Dehaene (2008), and likely did not encourage the use of benchmarking as much as it should. However, if this was true, it is telling that the response-grid model is not robust to task characteristics. For instance, several studies have found that when participants are given more time to perceive the stimulus (up to 6 s), multiple strategy use is the norm and that strategies employed depended on task characteristics (Gandini et al., 2010; Gandini, Lemaire, & Dufau, 2008; Luwel, Lemaire, & Verschaffel, 2005; Luwel, Verschaffel, Onghena, & De Corte, 2003). More recently, Cheyette and Piantadosi (2019) tracked eye movements when participants performed a numerosity estimation task with dot arrays presented from 100 to 3000 ms, and they found that participants do not process all the dots in each set in parallel, but via a serial accumulation of dots that are foveate within a given duration. The more time participants had to foveate, the more accurate their estimates were. Indeed, many of our participants reported using strategies that require serial accumulation across visual fixations (e.g., approximate or exact counting, and anchoring) as part of their repertoire. However, as acknowledged by the authors, their finding cannot distinguish whether "people build up an increasingly precise image of the visual scene as they saccade, from which numerical information is later extracted" (i.e., benchmarking), or that "numerical quantities themselves are being integrated across visual fixations" (i.e., non-benchmarking) (p. 5). Strategy reports can thus provide complementary information. In sum, our findings suggest that the response-grid account

at least in its current form is insufficient in explaining how numerical estimation is performed; even if a response grid exists, we believe that it does not remain stable throughout a task, and may undergo iterations of transformations depending on the strategy adopted on each trial. This could account for the discontinuity in the calibration effects. Regardless of which alternative is more likely, it is evident than the current assumptions of a response grid need to be updated to accommodate the discontinuity in calibration effects and how task characteristics may influence strategy choice. It is also clear that the distinction between associative and structural mappings is likely not a function of set size (although that is indeed the case when we consider only true null differences between calibration conditions), and that structural mapping does not exist in the continuous fashion previously thought. Finally, the current study provides the first evidence that the base range of associative mappings is neither related to estimation performance nor to math competence.

Alternative account of calibration effects observed globally rather than locally

How else can we account for the observations that calibration manipulations tend to affect not only the calibrated numerosity (locally), but other numerosities tested as well (globally) if we do not assume a holistic link between the analog mental number line and the symbolic response grid? We do not attempt to develop a full-fledged theory here based on the limits of our data, but we speculate that "structural mappings" may also exist at a more local level using analogical reasoning iteratively. Hence, the apparent calibration effects observed globally could arise from trial-to-trial dependencies from a Bayesian perspective (Cicchini, Anobile, & Burr, 2014; Petzschner, Glasauer, & Stephan, 2015). This hypothesis is not novel in the context of numerosity-numeral mappings and has previously been alluded to within the analogy-based account proposed by Sullivan and colleagues (Alvarez et al., 2017; Sullivan & Barner, 2013, 2014). For instance, mappings can be formed by inferential processes from prior mappings (e.g., if a set of 25 items was previously labeled as "20", a set of 50 items would be labeled as close to "40"), or using strong associative mappings with subsets as an anchor (e.g., if a set seems to be composed of about 5 groups of about 4 items each, then the set would be labeled as "20"). Such trial-to-trial dependencies have been demonstrated previously using a nonsymbolic number-line task in which participants were shown an array of dots and were asked to indicate its relative position on a line demarcated by two sample dot arrays (e.g., 1 and 100 dots) (Cicchini et al., 2014). Cicchini, Anobile, and Burr (2014) found that participants' responses on a trial strongly correlated with the numerosity on the immediately preceding trial, which suggests that the numerosity–space mappings are dynamically constructed rather than retrieved from static representations of number on a mental number line. The authors propose that taking into account recent history may reflect a general strategy that optimizes estimation behavior (Cicchini et al., 2014). Also using a number-line task, but with Arabic numerals instead of dot arrays, Sullivan, Juhasz, Slattery, and Barth (2011) found that participants' estimates were influenced by the numerical magnitude of the first number that they had to map its relative position to. Both of these dependence findings are consistent with why the base range of associative mappings may not matter as only a single numerosity-numeral mapping may be sufficient to constrain subsequent mappings.

Nonetheless, this trial-to-trial dependence account still cannot account for the discontinuity in the calibration effects. One possibility inferred from participants' strategy reports here and in previous studies (Gandini et al., 2010; Gandini, Lemaire, & Dufau, 2008; Luwel et al., 2005, 2003), as well as analyses of changes in estimation consistency upon calibration, is that even though participants may use numerosity-numeral mappings from previous trials to infer the mapping on the current trial, there are strategic changes from trial to trial that could disrupt any trial-to-trial dependencies. Specifically, the employment of a specific strategy depends partly on the set size and the configuration of the items in the set. Hence, from a participant's perspective, it may be optimal to construct the mappings dynamically by considering prior mappings and employing different strategies on different trials rather than rely on a single strategy.

Taken together, multiple strategies and trial-to-trial dependencies can account for both the discontinuity in calibration effects and the so-called structural mappings respectively, without the need to assume a system-level holistic mapping between an analog mental number line and a numeral response grid, or even those individual systems per se. In fact, based on the ad-hoc analogy-based account, it may be more appropriate to think of "structural mappings" as mappings that are dynamically manipulated from associative mappings. We propose that it is necessary for future studies to devise more direct methods to distinguish holistic "structural mappings" based on the response grid account and ad-hoc "structural mappings" based on the analogy-based and trial-to-trial dependence accounts.

Limitations

To minimize fatigue from completing hundreds of trials without a break, we had to strike a balance between the sampling range and the number of data points which we could collect for each numerosity. The compromise resulted in at most ten data points per numerosity between the uncalibrated and calibrated conditions, which might have reduced the precision in estimating each participant's base range

of associative mappings. The limited number of trials per numerosity could have led to non-significant calibration effects with inconclusive evidence, and, therefore, resulted in the observed discontinuities and an over-estimation of the true upper bound of the base range of associative mappings. However, analyses of the continuity of calibration effects using multiple approaches and measures suggest that the discontinuity was unlikely to be driven by statistical randomness, but meaningful variability such as strategy changes. Although low statistical power may cast doubts on our conclusion that the base range of associative mappings is not related to estimation performance or math competence, low power should largely result in a systematic over-estimation of the base range, which should not pose a significant problem for analyses of inter-individual differences. A related concern is that the lack of relation between the base range of associative mappings and the estimation performance and math measures could be due to low reliability of the measures considered. Although the novel measure of base range of associative mappings has slightly below acceptable split-half reliability, it has high consistency between different computational approaches. Future research may consider computing the base range index using a re-calibration paradigm (i.e., with an additional calibration condition with a very different calibration ratio, e.g., Izard & Dehaene, 2008). Nonetheless, even if such relations truly exist, our data also suggest that what may matter more is the total collection of associative mappings regardless of numerical size, rather than just the base range among small numbers.

Finally, although strategy reports can offer insights that inform theories, a delayed summative reporting format (about 40–60 min after completing both uncalibrated and calibrated conditions) might limit the specificity of data that we can potentially observe patterns in. Hence, we believe that a trial-level reporting format as employed by Gandini and colleagues (Gandini et al., 2010; Gandini, Lemaire, & Dufau, 2008) will allow future research to better construct a more comprehensive model of numerical estimation and calibration. Although skepticism about the validity of self-reports is understandable, there were no apparent aspects of the study protocol that could bias the reports, and participants were generally capable of introspection in their open-ended responses. Ultimately, self-reports are one of many measurement tools, and their validity depends on the research question (Haeffel & Howard, 2010). In investigations of arithmetic strategy use, at least, self-reports can even be superior to "objective" measures (Tschentscher & Hauk, 2014, 2015), or even necessary for making inferences from "objective" measures such as neuroimaging measures (Grabner et al., 2009; Matejko & Ansari, 2019; Peters & De Smedt, 2018; Polspoel, Peters, Vandermosten, & De Smedt, 2017).

It is also important to note that although the current sample have standardized scores in reading and math that span a wide range and are normally distributed, they were centered slightly above average (Ms = 118–124). Hence, our findings may be specific to an above-average sample. It may be plausible that the base range of associative mappings is related to estimation performance and math competence in an average or below-average sample, or even in younger populations in which the experience-based associative mappings are developing.

Conclusions

The quality of numerosity-numeral mappings during estimation seems robustly related to procedural calculation skills. We tested whether having a higher base range of associative numerosity-numeral mappings is related to estimation performance or math competence. It was not related to either, and is, therefore, unlikely to underlie the estimation–math relation. Critically, our data did not provide strong evidence for a holistic link between an analog mental number line and a symbolic response grid. We found discontinuities in calibration effects, suggesting either strong associative mappings among large numbers or weak structural links between the two systems. In either case, our results call into question the existence of "holistic structural mapping" (based on the response-grid model) as a reliable mechanism of transcoding between sets of objects and numerals. We also found that direct retrieval of numerosity-numeral mappings for large numbers is likely not an essential part of participants' strategy repertoires. Our findings, therefore, suggest that performing serial estimation tasks may involve taking into account existing or previously constructed numerosity-numeral mappings to further construct new mappings and flexibly switching among strategies (e.g., based on stimulus characteristics).

Acknowledgements We would like to thank Olivia Lasala, Nathaniel Dorris, Jordann Lewis, and Reginald Wimbley for their assistance with data collection. DJY is supported by the Humanities, Arts, and Social Sciences International PhD Scholarship, co-funded by Nanyang Technological University and the Ministry of Education (Singapore). We would also like to thank the reviewers for their insightful feedback and analytical suggestions on a prior version of this manuscript.

Funding This work was supported in part by a National Science Foundation Grant (DRL 1660816) awarded to G.R.P.

Data availability The data set used for the current study is available on Open Science Framework https://osf.io/tas82/.

Compliance with ethical standards

Conflict of interest We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Research involving human and animal participants All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent Informed consent was obtained from all individuals included in the study.

Appendix

Questionnaire

1. Did you notice anything unusual about the estimation tasks? [Yes/No]
    a. What was unusual about the tasks?
    b. How did this unusual factor affect your performance on the last task block?
2. How confident were you with your estimates on the first block as a whole, on a scale of 0 (not at all) - 100 (extremely)?
3. Which of the following strategies did you use to estimate each set of dots during the first task block? (check any that apply)
    a. No particular strategy that I was aware of
    b. Exact counting (e.g., "I counted exactly in groups of X, and I am certain there are X dots in each group.")
        i. What was the largest quantity that you could label with absolute confidence without counting?
    c. Approximate counting (e.g., "I first saw one group of about X dots, a group of about Y dots, and another group of about Z dots. So, there were about (X + Y + Z) dots.")
    d. Instinctively knew each and every quantity from memory
    e. Retrieved a quantity from memory, and approximately added or subtracted from it (e.g., "I quickly looked at all the dots, thought it looked like there are about X or a little bit more, so I said slightly more than X.")
        i. Are there particular quantities that you retrieved from memory?
            1. Yes
                a. What particular quantity (or quantities) did you retrieve from memory? Please list them.
            2. No, I usually used the trial before as a reference quantity for the current trial
    f. Used a subset of the dots as an anchor (e.g., "I counted exactly a group of X dots, or saw a group about X dots. Then, I estimated that there were six other similar groups, so I figured that there are 7X dots.")
        i. Are there particular quantities that you used as an anchor set/chunk?
            1. Yes
                a. What particular quantity (or quantities) did you use as an anchor set/chunk? Please list them.
            2. No, I usually used the trial before as an anchor set/chunk for the current trial
    g. Others
        i. Please elaborate on what strategy or strategies you used.
4. You mentioned that you used more than one strategy. Under what situations did you use each strategy (i.e. particular types of trials, quantities, time into the task block, etc.)?
5. Did your strategies differ between the first block and the last block of trials? [Yes/No]
    a. How did your strategies differ?
6. How confident were you with your estimates on the last block as a whole, on a scale of 0 (not at all) - 100 (extremely)?

References

Alvarez, J., Abdul-Chani, M., Deutchman, P., DiBiasie, K., Iannucci, J., Lipstein, R., et al. (2017). Estimation as analogy-making: Evidence that preschoolers' analogical reasoning ability predicts their numerical estimation. *Cognitive Development, 41*, 73–84. https://doi.org/10.1016/j.cogdev.2016.12.004.

Bartelet, D., Vaessen, A., Blomert, L., & Ansari, D. (2014). What basic number processing measures in kindergarten explain unique variability in first-grade arithmetic proficiency? *Journal of Experimental Child Psychology, 117*(1), 12–28. https://doi.org/10.1016/j.jecp.2013.08.010.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*. https://doi.org/10.2307/2346101.

Bishara, A. J., & Hittner, J. B. (2012). Testing the significance of a correlation with nonnormal data: Comparison of Pearson, Spearman, transformation, and resampling approaches. *Psychological Methods, 17*(3), 399–417. https://doi.org/10.1037/a0028087.

Bishara, A. J., & Hittner, J. B. (2015). Reducing bias and error in the correlation coefficient due to nonnormality. *Educational and Psychological Measurement, 75*(5), 785–804. https://doi.org/10.1177/0013164414557639.

Booth, J. L., & Siegler, R. S. (2006). Developmental and individual differences in pure numerical estimation. *Developmental Psychology, 42*(1), 189–201. https://doi.org/10.1037/0012-1649.41.6.189.

Brankaer, C., Ghesquière, P., & De Smedt, B. (2014). Children's mapping between non-symbolic and symbolic numerical magnitudes and its association with timed and untimed tests of mathematics achievement. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0093565.

Burr, D. C., Turi, M., & Anobile, G. (2010). Subitizing but not estimation of numerosity requires attentional resources. *Journal of Vision, 10*(2010), 1–10. https://doi.org/10.1167/10.6.20.

Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. New York, NY: Routledge. https://doi.org/10.4324/9781410600219.

Carey, S., & Barner, D. (2019). Ontogenetic origins of human integer representations. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2019.07.004.

Carey, S., Shusterman, A., Haward, P., & Distefano, R. (2017). Do analog number representations underlie the meanings of young children's verbal numerals? *Cognition, 168*, 243–255. https://doi.org/10.1016/j.cognition.2017.06.022.

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

Chesney, D. L., Bjalkebring, P., & Peters, E. (2015). How to estimate how well people estimate: Evaluating measures of individual differences in the approximate number system. *Attention, Perception, & Psychophysics, 77*(8), 2781–2802. https://doi.org/10.3758/s13414-015-0974-6.

Cheyette, S. J., & Piantadosi, S. T. (2019). A primarily serial, foveal accumulator underlies approximate numerical estimation. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1819956116.

Cicchini, G. M., Anobile, G., & Burr, D. C. (2014). Compressive mapping of number to space reflects dynamic encoding mechanisms, not static logarithmic transform. *Proceedings of the National Academy of Sciences, 111*(21), 7867–7872. https://doi.org/10.1073/pnas.1402785111.

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review, 12*(5), 769–786. https://doi.org/10.3758/BF03196772.

Cutini, S., Scatturin, P., Basso Moro, S., & Zorzi, M. (2014). Are the neural correlates of subitizing and estimation dissociable? An fNIRS investigation. *NeuroImage, 85*(1, SI), 391–399. https://doi.org/10.1016/j.neuroimage.2013.08.027.

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use welch's *t*-test instead of student's *t*-test. *International Review of Social Psychology*, *30*(1), 92. https://doi.org/10.5334/irsp.82.

Dehaene, S. (2007). Symbols and quantities in parietal cortex: Elements of a mathematical theory of number representation and manipulation. In P. Haggard, Y. Rossetti, & M. Kawato (Eds.), *Attention & performance XXII. Sensorimotor Foundations of Higher Cognition* (pp. 527–574). Cambridge, MA: Harvard University Press. https://doi.org/10.1093/acprof:oso/9780199231447.003.0024.

Dehaene, S., & Changeux, J. P. (1993). Development of elementary numerical abilities: A neuronal model. *Journal of Cognitive Neuroscience, 5*(4), 390–407. https://doi.org/10.1162/jocn.1993.5.4.390.

Dehaene, S., Izard, V., & Piazza, M. (2005). *Control over non-numerical parameters in numerosity experiments*. Retrieved from www.unicog.org/docs/DocumentationDotsGeneration.doc.

Dehaene, S., & Mehler, J. (1992). Cross-linguistic regularities in the frequency of number words. *Cognition, 43*(1), 1–29. https://doi.org/10.1016/0010-0277(92)90030-L.

Dehaene, S., Spelke, E. S., Pinel, P., Stanescu, R., & Tsivkin, S. (1999). Sources of mathematical thinking: Behavioral and brain-imaging evidence. *Science, 284*(5416), 970–974. https://doi.org/10.1126/science.284.5416.970.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*(July), 1–17. https://doi.org/10.3389/fpsyg.2014.00781.

Ebersbach, M., & Erz, P. (2014). Symbolic versus non-symbolic magnitude estimations among children and adults. *Journal of Experimental Child Psychology, 128*, 52–68. https://doi.org/10.1016/j.jecp.2014.06.005.

Foster, J. L., Shipstead, Z., Harrison, T. L., Hicks, K. L., Redick, T. S., & Engle, R. W. (2015). Shortened complex span tasks can reliably measure working memory capacity. *Memory & Cognition, 43*(2), 226–236. https://doi.org/10.3758/s13421-014-0461-7.

Fox, J., & Weisberg, S. (2011). *An {R} companion to applied regression* (Second). Thousand Oaks, CA: Sage. Retrieved from https://socserv.socsci.mcmaster.ca/jfox/Books/Companion.

Gandini, D., Ardiale, E., & Lemaire, P. (2010). Children's strategies in approximate quantification. *Current Psychology Letters*, *26*(1). Retrieved from https://cpl.revues.org/4990.

Gandini, D., Lemaire, P., Anton, J. L., & Nazarian, B. (2008). Neural correlates of approximate quantification strategies in young and older adults: An fMRI study. *Brain Research, 1246*, 144–157. https://doi.org/10.1016/j.brainres.2008.09.096.

Gandini, D., Lemaire, P., & Dufau, S. (2008). Older and younger adults' strategies in approximate quantification. *Acta Psychologica, 129*(1), 175–189. https://doi.org/10.1016/j.actpsy.2008.05.009.

Gonthier, C., Thomassin, N., & Roulin, J. L. (2016). The composite complex span: French validation of a short working memory task. *Behavior Research Methods, 48*(1), 233–242. https://doi.org/10.3758/s13428-015-0566-3.

Grabner, R. H., Ansari, D., Koschutnig, K., Reishofer, G., Ebner, F., & Neuper, C. (2009). To retrieve or to calculate? Left angular gyrus mediates the retrieval of arithmetic facts during problem solving. *Neuropsychologia, 47*(2), 604–608. https://doi.org/10.1016/j.neuropsychologia.2008.10.013.

Guillaume, M., Gevers, W., & Content, A. (2016). Assessing the approximate number system: No relation between numerical

comparison and estimation tasks. *Psychological Research Psychologische Forschung, 80*(2), 248–258. https://doi.org/10.1007/s00426-015-0657-x.

Haeffel, G. J., & Howard, G. S. (2010). Self-report: Psychology's four-letter word. *The American Journal of Psychology, 123*(2), 181. https://doi.org/10.5406/amerjpsyc.123.2.0181.

Hair, J. F., Jr., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Upper Saddle River, New Jersey: Prentice Hall.

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics, 42*(2), 377–381. https://doi.org/10.1016/j.jbi.2008.08.010.

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.

Hyde, D. C. (2011). Two systems of non-symbolic numerical cognition. *Frontiers in Human Neuroscience*. https://doi.org/10.3389/fnhum.2011.00150.

Izard, V., & Dehaene, S. (2008). Calibrating the mental number line. *Cognition, 106*(3), 1221–1247. https://doi.org/10.1016/j.cognition.2007.06.004.

Jang, S., & Cho, S. (2016). The acuity for numerosity (but not continuous magnitude) discrimination correlates with quantitative problem solving but not routinized arithmetic. *Current Psychology, 35*(1), 44–56. https://doi.org/10.1007/s12144-015-9354-6.

Jang, S., & Cho, S. (2018). The mediating role of number-to-magnitude mapping precision in the relationship between approximate number sense and math achievement depends on the domain of mathematics and age. *Learning and Individual Differences, 64*(May 2017), 113–124. https://doi.org/10.1016/j.lindif.2018.05.005.

JASP Team. (2019). JASP (Version 0.9.2.0) [Computer software]. Retrieved from https://jasp-stats.org/.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *The American Journal of Psychology, 62*(4), 498. https://doi.org/10.2307/1418556.

Kline, R. B. (2011). *Principles and practice of structural equation modeling*. New York, NY: The Guilford Press. Retrieved from http://www.guilford.com/companion-site/Principles-and-Practice-of-Structural-Equation-Modeling-Third-Edition%5Cnhttp://www.guilford.com/books/Principles-and-Practice-of-Structural-Equation-Modeling/Rex-B-Kline/9781606238769%5Cnhttp://www.psych.umass.edu/u

Krueger, L. E. (1984). Perceived numerosity: A comparison of magnitude production, magnitude estimation, and discrimination judgments. *Perception & Psychophysics, 35*(6), 536–542. https://doi.org/10.3758/BF03205949.

Langsrud, Ø. (2003). ANOVA for unbalanced data: Use Type II instead of Type III sums of squares. *Statistics and Computing, 13*(2), 163–167. https://doi.org/10.1023/A:1023260610025.

Le Corre, M., & Carey, S. (2007). One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition, 105*(2), 395–438. https://doi.org/10.1016/j.cognition.2006.10.005.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139087759.

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49*(4), 764–766. https://doi.org/10.1016/j.jesp.2013.03.013.

Libertus, M. E., Feigenson, L., & Halberda, J. (2013). Numerical approximation abilities correlate with and predict informal but not formal mathematics abilities. *Journal of Experimental Child Psychology, 116*(4), 829–838. https://doi.org/10.1016/j.jecp.2013.08.003.

Libertus, M. E., Feigenson, L., Halberda, J., & Landau, B. (2014). Understanding the mapping between numerical approximation and number words: Evidence from Williams syndrome and typical development. *Developmental Science, 17*(6), 905–919. https://doi.org/10.1111/desc.12154.

Libertus, M. E., Odic, D., Feigenson, L., & Halberda, J. (2016). The precision of mapping between number words and the approximate number system predicts children's formal math abilities. *Journal of Experimental Child Psychology, 150*, 207–226. https://doi.org/10.1016/j.jecp.2016.06.003.

Lipton, J. S., & Spelke, E. S. (2005). Preschool children's mapping of number words to nonsymbolic numerosities. *Child Development, 76*(5), 978–988. https://doi.org/10.1111/j.1467-8624.2005.00891.x.

Lourenco, S. F., Bonny, J. W., Fernandez, E. P., & Rao, S. (2012). Nonsymbolic number and cumulative area representations contribute shared and unique variance to symbolic math competence. *Proceedings of the National Academy of Sciences, 109*(46), 18737–18742. https://doi.org/10.1073/pnas.1207212109.

Luwel, K., Lemaire, P., & Verschaffel, L. (2005). Children's strategies in numerosity judgment. *Cognitive Development, 20*(3), 448–471. https://doi.org/10.1016/j.cogdev.2005.05.007.

Luwel, K., Verschaffel, L., Onghena, P., & De Corte, E. (2003). Strategic aspects of numerosity judgment: The effect of task characteristics. *Experimental Psychology, 50*(1), 63–75. https://doi.org/10.1026//1618-3169.50.1.63.

Lyons, I. M., & Ansari, D. (2009). The cerebral basis of mapping nonsymbolic numerical quantities onto abstract symbols: An fMRI training study. *Journal of Cognitive Neuroscience, 21*(9), 1720–1735. https://doi.org/10.1162/jocn.2009.21124.

Lyons, I. M., & Beilock, S. L. (2009). Beyond quantity: Individual differences in working memory and the ordinal understanding of numerical symbols. *Cognition, 113*(2), 189–204. https://doi.org/10.1016/j.cognition.2009.08.003.

Lyons, I. M., Bugden, S., Zheng, S., De Jesus, S., & Ansari, D. (2018). Symbolic number skills predict growth in nonsymbolic number skills in kindergarteners. *Developmental Psychology, 54*(3), 440–457. https://doi.org/10.1037/dev0000445.

Lyons, I. M., Price, G. R., Vaessen, A., Blomert, L., & Ansari, D. (2014). Numerical predictors of arithmetic success in grades 1–6. *Developmental Science, 17*(5), 714–726. https://doi.org/10.1111/desc.12152.

Malone, S. A., Heron-delaney, M., Burgoyne, K., & Hulme, C. (2019). Learning correspondences between magnitudes, symbols and words: Evidence for a triple code model of arithmetic development. *Cognition, 187*(March 2018), 1–9. https://doi.org/10.1016/j.cognition.2018.11.016.

Matejko, A. A., & Ansari, D. (2019). The neural association between arithmetic and basic numerical processing depends on arithmetic problem size and not chronological age. *Developmental Cognitive Neuroscience, 37*(April), 100653. https://doi.org/10.1016/j.dcn.2019.100653.

Mazzocco, M. M. M., Feigenson, L., & Halberda, J. (2011a). Impaired acuity of the approximate number system underlies mathematical learning disability (dyscalculia). *Child Development, 82*(4), 1224–1237. https://doi.org/10.1111/j.1467-8624.2011.01608.x.

Mazzocco, M. M. M., Feigenson, L., & Halberda, J. (2011b). Preschoolers' precision of the approximate number system predicts

later school mathematics performance. *PLoS ONE, 6*(9), e23749. https://doi.org/10.1371/journal.pone.0023749.

Mejias, S., Grégoire, J., & Noël, M.-P. (2012a). Numerical estimation in adults with and without developmental dyscalculia. *Learning and Individual Differences, 22*(1), 164–170. https://doi.org/10.1016/j.lindif.2011.09.013.

Mejias, S., Mussolin, C., Rousselle, L., Grégoire, J., & Noël, M.-P. (2012b). Numerical and nonnumerical estimation in children with and without mathematical learning disabilities. *Child Neuropsychology, 18*(6), 550–575. https://doi.org/10.1080/09297049.2011.625355.

Mejias, S., & Schiltz, C. (2013). Estimation abilities of large numerosities in Kindergartners. *Frontiers in Psychology, 4*(August), 518. https://doi.org/10.3389/fpsyg.2013.00518.

Merkley, R., & Scerif, G. (2015). Continuous visual properties of number influence the formation of novel symbolic representations. *Quarterly Journal of Experimental Psychology, 68*(9), 1860–1870. https://doi.org/10.1080/17470218.2014.994538.

Merkley, R., Shimi, A., & Scerif, G. (2016). Electrophysiological markers of newly acquired symbolic numerical representations: The role of magnitude and ordinal information. *ZDM, 48*(3), 279–289. https://doi.org/10.1007/s11858-015-0751-y.

Minturn, A. L., & Reese, T. W. (1951). The effect of differential reinforcement on the discrimination of visual number. *The Journal of Psychology, 31*(2), 201–231. https://doi.org/10.1080/00223980.1951.9712804.

Mundy, E., & Gilmore, C. K. (2009). Children's mapping between symbolic and nonsymbolic representations of number. *Journal of Experimental Child Psychology, 103*(4), 490–502. https://doi.org/10.1016/j.jecp.2009.02.003.

Mussolin, C., Nys, J., Content, A., & Leybaert, J. (2014). Symbolic number abilities predict later approximate number system acuity in preschool children. *PLoS ONE, 9*(3), e91839. https://doi.org/10.1371/journal.pone.0091839.

Opfer, J. E., & Siegler, R. S. (2007). Representational change and children's numerical estimation. *Cognitive Psychology, 55*(3), 169–195. https://doi.org/10.1016/j.cogpsych.2006.09.002.

Orrantia, J., Muñez, D., Matilla, L., Sanchez, R., San Romualdo, S., & Verschaffel, L. (2019). Disentangling the mechanisms of symbolic number processing in adults' mathematics and arithmetic achievement. *Cognitive Science, 43*(1), 1–24. https://doi.org/10.1111/cogs.12711.

Peters, L., & De Smedt, B. (2018). Arithmetic in the developing brain: A review of brain imaging studies. *Developmental Cognitive Neuroscience, 30*(May 2017), 265–279. https://doi.org/10.1016/j.dcn.2017.05.002.

Petzschner, F. H., Glasauer, S., & Stephan, K. E. (2015). A Bayesian perspective on magnitude estimation. *Trends in Cognitive Sciences, 19*(5), 285–293. https://doi.org/10.1016/j.tics.2015.03.002.

Piazza, M., Fumarola, A., Chinello, A., & Melcher, D. (2011). Subitizing reflects visuo-spatial object individuation capacity. *Cognition, 121*(1), 147–153. https://doi.org/10.1016/j.cognition.2011.05.007.

Piazza, M., Pinel, P., Le Bihan, D., & Dehaene, S. (2007). A magnitude code common to numerosities and number symbols in human intraparietal cortex. *Neuron, 53*(2), 293–305. https://doi.org/10.1016/j.neuron.2006.11.022.

Pincham, H. L., & Szucs, D. (2012). Intentional subitizing: Exploring the role of automaticity in enumeration. *Cognition, 124*(2), 107–116. https://doi.org/10.1016/j.cognition.2012.05.010.

Pinheiro-Chagas, P., Dotan, D., Piazza, M., & Dehaene, S. (2017). Finger tracking reveals the covert stages of mental arithmetic. *Open Mind, 1*(1), 30–41. https://doi.org/10.1162/OPMI_a_00003.

Pinheiro-Chagas, P., Wood, G., Knops, A., Krinzinger, H., Lonnemann, J., Starling-Alves, I., et al. (2014). In how many ways is the approximate number system associated with exact calculation? *PLoS ONE, 9*(11), e111155. https://doi.org/10.1371/journal.pone.0111155.

Polspoel, B., Peters, L., Vandermosten, M., & De Smedt, B. (2017). Strategy over operation: neural activation in subtraction and multiplication during fact retrieval and procedural strategy use in children. *Human Brain Mapping, 38*(9), 4657–4670. https://doi.org/10.1002/hbm.23691.

Price, J., Clement, L. M., & Wright, B. J. (2014). The role of feedback and dot presentation format in younger and older adults' number estimation. *Aging, Neuropsychology, and Cognition, 21*(1), 68–98. https://doi.org/10.1080/13825585.2013.786015.

Revkin, S. K., Piazza, M., Izard, V., Cohen, L., & Dehaene, S. (2008). Does subitizing reflect numerical estimation? *Psychological Science, 19*(6), 607–614. https://doi.org/10.1111/j.1467-9280.2008.02130.x.

Starr, A., Libertus, M. E., & Brannon, E. M. (2013). Number sense in infancy predicts mathematical abilities in childhood. *Proceedings of the National Academy of Sciences of the United States of America, 110*(45), 18116–18120. https://doi.org/10.1073/pnas.1302751110.

Stoianov, I. (2014). Generative processing underlies the mutual enhancement of arithmetic fluency and math-grounding number sense. *Frontiers in Psychology, 5*(Nov), 1–4. https://doi.org/10.3389/fpsyg.2014.01326.

Stoianov, I., & Zorzi, M. (2012). Emergence of a "visual number sense" in hierarchical generative models. *Nature Neuroscience, 15*(2), 194–196. https://doi.org/10.1038/nn.2996.

Suárez-Pellicioni, M., & Booth, J. R. (2018). Fluency in symbolic arithmetic refines the approximate number system in parietal cortex. *Human Brain Mapping, 39*(10), 3956–3971. https://doi.org/10.1002/hbm.24223.

Sullivan, J., & Barner, D. (2013). How are number words mapped to approximate magnitudes? *The Quarterly Journal of Experimental Psychology, 66*(2), 389–402. https://doi.org/10.1080/17470218.2012.715655.

Sullivan, J., & Barner, D. (2014). Inference and association in children's early numerical estimation. *Child Development, 85*(4), 1740–1755. https://doi.org/10.1111/cdev.12211.

Sullivan, J., Frank, M. C., & Barner, D. (2016). Intensive math training does not affect approximate number acuity: Evidence from a three-year longitudinal curriculum intervention. *Journal of Numerical Cognition, 2*(2), 57–76. https://doi.org/10.5964/jnc.v2i2.19.

Sullivan, J., Juhasz, B. J., Slattery, T. J., & Barth, H. C. (2011). Adults' number-line estimation strategies: Evidence from eye movements. *Psychonomic Bulletin & Review, 18*(3), 557–563. https://doi.org/10.3758/s13423-011-0081-1.

The jamovi project. (2019). jamovi (Version 0.9) [Computer software]. Retrieved from https://www.jamovi.org/.

Trick, L. M., & Pylyshyn, Z. W. (1994). Why are small and large numbers enumerated differently? A limited-capacity preattentive stage in vision. *Psychological Review, 101*(1), 80–102. https://doi.org/10.1037/0033-295X.101.1.80.

Tschentscher, N., & Hauk, O. (2014). How are things adding up? Neural differences between arithmetic operations are due to general problem solving strategies. *NeuroImage, 92*, 369–380. https://doi.org/10.1016/j.neuroimage.2014.01.061.

Tschentscher, N., & Hauk, O. (2015). Individual strategy ratings improve the control for task difficulty effects in arithmetic problem solving paradigms. *Frontiers in Psychology, 6*, 1188. https://doi.org/10.3389/fpsyg.2015.01188.

Verguts, T., & Fias, W. (2004). Representation of number in animals and humans: A neural model. *Journal of Cognitive Neuroscience, 16*(9), 1493–1504. https://doi.org/10.1162/0898929042568497.

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., et al. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review, 25*(1), 58–76. https://doi.org/10.3758/s13423-017-1323-7.

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., et al. (2017). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin and Review*. https://doi.org/10.3758/s13423-017-1343-3.

Wong, T. T.-Y., Ho, C. S.-H., & Tang, J. (2016a). Consistency of response patterns in different estimation tasks. *Journal of Cognition and Development, 17*(3), 526–547. https://doi.org/10.1080/15248372.2015.1072091.

Wong, T. T.-Y., Ho, C. S.-H., & Tang, J. (2016b). The relation between ANS and symbolic arithmetic skills: The mediating role of number-numerosity mappings. *Contemporary Educational Psychology, 46*, 208–217. https://doi.org/10.1016/j.cedpsych.2016.06.003.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock–Johnson III tests of achievement*. Itasca, IL: Riverside.

Yeo, D. J., Wilkey, E. D., & Price, G. R. (2019). Malleability of mappings between Arabic numerals and approximate quantities: Factors underlying individual differences and the relation to math. *Acta Psychologica, 198*(June), 102877. https://doi.org/10.1016/j.actpsy.2019.102877.

Zeileis, A., & Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News*, *2*(3), 7–10. Retrieved from https://cran.r-project.org/doc/Rnews/.

Zhao, H., Chen, C., Zhang, H., Zhou, X., Mei, L., Chen, C., et al. (2012). Is order the defining feature of magnitude representation? An ERP study on learning numerical magnitude and spatial order of artificial symbols. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0049565.

Zimmerman, D. W., & Zumbo, B. D. (1993). Rank transformations and the power of the student T-test and Welch T-test for non-normal populations with unequal variances. *Canadian Journal of Experimental Psychology, 47*(3), 523–539. https://doi.org/10.1037/h0078850.