Acta Psychologica 198 (2019) 102877

Contents lists available at ScienceDirect

Acta Psychologica

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

Malleability of mappings between Arabic numerals and approximate quantities: Factors underlying individual differences and the relation to math

Darren J. Yeo[a,b], Eric D. Wilkey[a], Gavin R. Price[a,*]

[a] Department of Psychology & Human Development, Peabody College, Vanderbilt University, 230 Appleton Place, Nashville, TN 37203, United States of America
[b] Division of Psychology, School of Social Sciences, Nanyang Technological University, 48 Nanyang Avenue, Singapore 639818, Singapore

ARTICLE INFO

Keywords:
Numerical cognition
Numerosity estimation
Calibration
Mathematical competence
Individual differences

ABSTRACT

Humans tend to be inaccurate and inconsistent when estimating a large number of objects. Furthermore, we modify our estimates when feedback or a reference array is provided, indicating that the mappings between perceived numerosity and their corresponding numerals are largely malleable in response to calibration. However, there is great variability in response to calibration across individuals. Using uncalibrated and calibrated numerosity estimation conditions, the current study explored the factors underlying individual differences in the extent and nature of the malleability of numerosity estimation performance as a result of calibration in a sample of 71 undergraduate students. We found that individual differences in performance were reliable across conditions, and participants' responses to calibration varied greatly. Participants who were less consistent or had more proportionally spaced (i.e., linear) estimates before calibration tended to shift the distributions of their estimates to a greater extent. Higher calculation competence also predicted an increase in how linear participants' estimates were after calibration. Moreover, the effect of calibration was not continuous across numerosities within participants. This suggests that the mechanisms underlying numeral-numerosity mappings may be less systematic than previously thought and likely depend on cognitive mechanisms beyond representation of numerosities. Taken together, the mappings between numerosities and numerical symbols may not be stable and direct, but transient and mediated by task-related (e.g., strategic) mechanisms. Rather than estimation skills being foundational for math competence, math competence may also influence estimation skills. Therefore, numerosity estimation tasks are not a pure measure of number representations.

Magnitude estimation is an essential skill in a variety of contexts, including computation (e.g., how much to tip for service), measurement (e.g., how far away), and numerosity (i.e., how many) judgments (Booth & Siegler, 2006; Hogan & Brezinski, 2003). However, adults who have experience with a wide range of numerical and non-numerical magnitudes tend to be very inaccurate even on the most basic of these skills – numerosity estimation. Specifically, in a typical numerosity estimation task (e.g., providing an estimate of how many dots there are in an array; hereafter "numerosity-to-numeral mappings"), individuals tend to systematically underestimate, particularly for relatively large numerosities (e.g., Crollen, Castronovo, & Seron, 2011; Izard & Dehaene, 2008; Krueger, 1984; Minturn & Reese, 1951) or magnitudes in general such as mass (Stevens, 1957). In addition to being inaccurate, responses between and within individuals are often highly variable (Cordes, Gelman, Gallistel, & Whalen, 2001; Izard & Dehaene, 2008; Whalen, Gallistel, & Gelman, 1999). However, estimation skills can be improved when individuals are provided with calibration for their estimates, providing a window into the processes involved in numerical estimation. The current study investigates the factors underlying individual differences in the extent and nature of the malleability of numerosity estimation performance as a result of calibration.

1. Local-to-global calibration

Several studies have demonstrated that when children and adults are given some form of calibration before a task (e.g., showing 30 dots and labeling it as "30") or feedback during the actual task (e.g., providing the correct number of dots after an estimate is made), participants modify their subsequent estimates, suggesting that the mappings between numerals and numerosities are malleable (Izard & Dehaene, 2008; Krueger, 1984; Price, Clement, & Wright, 2014). To further

* Corresponding author.
E-mail address: gavin.price@vanderbilt.edu (G.R. Price).

https://doi.org/10.1016/j.actpsy.2019.102877
Received 17 January 2019; Received in revised form 26 April 2019; Accepted 25 June 2019
0001-6918/ © 2019 Elsevier B.V. All rights reserved.

investigate the nature and extent of the malleability of estimation performance, researchers have begun exploring the use of misleading calibration. These studies found that accuracy of the calibration did not matter as even adults lack certainty about how large each set actually is (Izard & Dehaene, 2008; Sullivan & Barner, 2013). Hence, by employing different calibration values within the same participants, mechanisms underlying the numerosity-to-numeral mappings have been proposed (Izard & Dehaene, 2008). For example, Izard and Dehaene (2008) had adults estimate the numerosity of a series of dot arrays, after being presented with an explicitly mislabeled set of dots as an inducer (e.g., 25 dots labeled as "30" being an "overestimating inducer", or 39 dots labeled as "30" being an "underestimating inducer"). Relative to the veridical numerosity-to-numeral mappings, they found that participants' calibrated estimates were in the direction of the inducer (i.e., larger for the "overestimating inducer", and smaller for the "underestimating inducer"), not only locally for the calibrated numerosity (30) but also extended to the entire range of numerosities (9–100) tested. This phenomenon in which information about a *single* numerosity extends to the *entire range* of numerosities tested will be referred to here as "local-to-global" calibration, and suggests that numerosity-to-numeral mappings are not only malleable, but are also highly interdependent.

Izard and Dehaene (2008) postulated that mechanisms underlying local-to-global calibration for large numerosities may result from a mixture of unintentional "automatic learning process" (p. 1234) and conscious, strategic modification that may involve approximate arithmetic (as their participants had indicated via self-report). They proposed a response-grid model of numerosity estimation that comprises two stages (Fig. 1): (a) Encoding stage: A logarithmically scaled "mental number line" is divided into several segments, each of which corresponds to a different verbal label (e.g., 10's, 20's, 30's …), defining a *veridical* "response grid". This response grid thus serves as an interface between the analog mental number line and the symbolic number system. An encoded numerosity activates a point on the number line, which is translated into a verbal label; (b) Response selection stage: Individuals rarely possess a veridical response grid, but an idiosyncratic affine-transformed (i.e., stretched or compressed, and/or shifted globally) version of it. In the absence of an external calibration, a *spontaneous* (i.e., internally or self-calibrated) response grid is employed. In the presence of an external calibration, an *externally calibrated* response grid is employed. These transformations may or may not be conscious and strategic in either scenario (for more details of the model, see Izard & Dehaene, 2008). Crucially, their data also suggest that the calibration likely takes place during the process of response selection (i.e., via a bias in symbolic labelling), rather than during the perceptual encoding and discrimination of numerosities (Izard & Dehaene, 2008). The relatively stable discrimination sensitivity is a potential mechanism for constraining an individual's estimation performance across conditions.

However, small numerosities are less influenced by calibration than large numerosities (Alvarez et al., 2017; Sullivan & Barner, 2013, 2014). Sullivan and Barner (2013) had participants first complete an uncalibrated estimation task assessing numerosities sparsely sampled from 8 to 350. Then, they told participants that the largest set of dots they would see in a subsequent estimation task was 75, 375 or 750 when in fact the largest set they saw was still 350. Their results showed that the misleading upper bounds (a local calibration) induced a global shift in participants' estimates across the range of numerosities (8–350) tested in the direction of the inducers in a majority of their participants, replicating Izard and Dehaene's (2008) findings even with mere suggestions of an upper bound for the to-be-estimated numerosities. Yet, the calibration did not have as much of an impact on numerosities 8 and 12 as on numerosities larger than or equal to 20 (Sullivan & Barner, 2013). This led them to propose a distinction between small and large numerosities in the way symbolic and nonsymbolic representations of numerosity are mapped. Specifically, they suggest that numerosity-to-numeral mappings generally occur at a system level due to their analogous ordinal structures, rendering the mappings interdependent or "structurally mapped", consistent with the response-grid model. However, small numerosities tend to be more "associatively mapped" to numerals as a result of experience (Dehaene & Mehler, 1992; Lipton & Spelke, 2005; Verguts & Fias, 2004) and are more reliably and independently mapped to their true symbolic labels than larger numerosities. Hence, small numerosities were observed to be more resistant to calibration than larger numerosities (Fig. 1). Sullivan and Barner (2013) do not view "associative mapping" and "structural mapping" as mutually exclusive for a particular numerosity, and it is the relative strengths of each type of mapping that determines the susceptibility of a

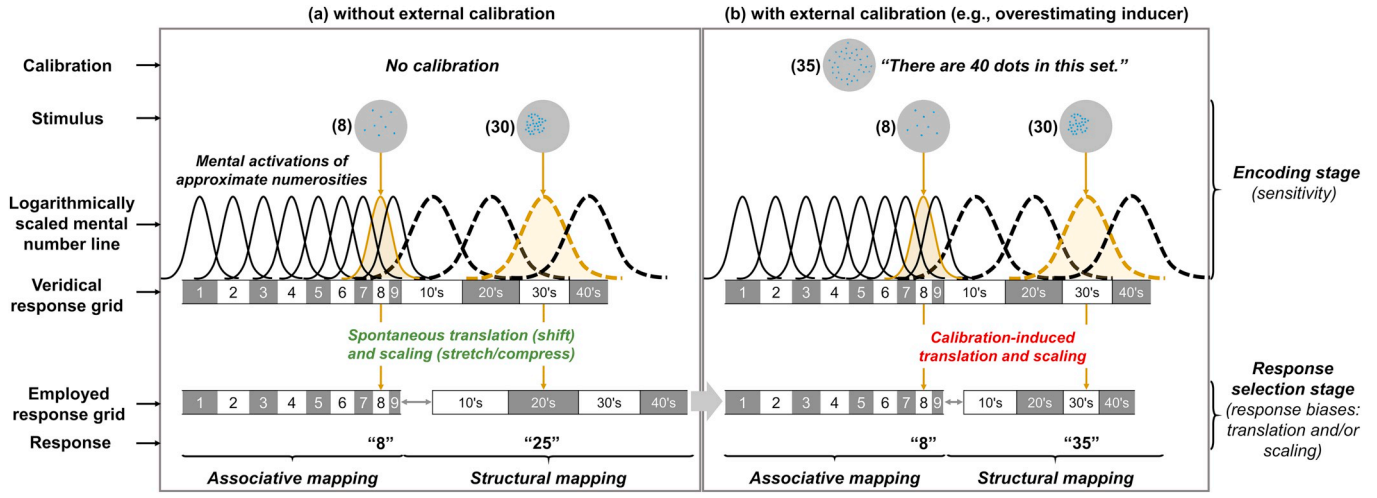

Fig. 1. A summary schematic of the response-grid model of numerosity estimation proposed by Izard and Dehaene (2008) and the associative mapping – structural mapping distinction proposed by Sullivan and Barner (2013), (a) without and (b) with an external calibration. The mental activations modeled as Gaussian curves are shown individually (solid curves) for the associative mapping range, and averaged across a range of possible numerosities (dotted curves) that could be mapped to a particular segment (e.g., 10's, 20's, 30's) of the response grid for the structural mapping range (e.g., see Revkin, Piazza, Izard, Cohen, & Dehaene, 2008). The employed response grids are idiosyncratic, and the ones illustrated here are for a hypothetical individual. Without an external calibration, participants have a general tendency to underestimate larger quantities (e.g., an array of 30 dots as "25"). Calibration likely takes place during the response selection stage rather than the encoding stage. Small numerosities (e.g., 8) are hypothesized to be more resistant to an internally (spontaneous) or externally induced (e.g., an array of 35 dots associated with "40") calibration than large numerosities (e.g., 30) are.
Adapted from Izard and Dehaene (*2008*), Calibrating the mental number line, Cognition, 106(3), 1221–1247, with permission from Elsevier.

particular numerosity to calibration. Using the same paradigm, this associative mapping–structural mapping distinction has been replicated in 4- to 7-year-old children (Alvarez et al., 2017; Sullivan & Barner, 2014).

In sum, previous research suggests that while small numerosities appear to be relatively resistant to calibration, larger numerosities are subject to local-to-global calibration effects and such effects seem robust across paradigms and age groups.

2. Individual differences in local-to-global calibration

Despite the studies above demonstrating that calibration has robust global effects in both children and adults at the group level, there are large individual differences in participants' responses to calibration. For instance, Sullivan and Barner (2013) reported that response to calibration was observed in 70–90% of adults. Similarly, other studies that used the same paradigm reported calibration effects in only 47–60% of 4- to 7-year-old children (Alvarez et al., 2017; Sullivan & Barner, 2014). Moreover, the smallest numerosity affected by calibration varied greatly across adults (ranging from 8 to 180; Sullivan & Barner, 2013) and 5- to 7-year-old children (ranging from 6 to 32; Sullivan & Barner, 2014). Using the smallest numerosity affected by calibration at the individual level as a metric of estimation malleability assumes that subsequent numerosities are also affected by calibration, but there is as yet no evidence supporting such continuity. It is possible that there exist relatively strong associative mappings for larger numerosities at the individual level that may be acquired through experience and may serve as idiosyncratic anchors (e.g., Dehaene & Mehler, 1992). Nonetheless, existing findings suggest large individual differences in the relative strengths of associative versus structural mapping across numerosities, and that the distinction between associative mapping and structural mapping may not be purely a function of set size.

Furthermore, individual differences in the calibration effect may be sensitive to the order of calibration. In Izard and Dehaene's (2008) study, every participant underwent two sessions of estimation, each with a different inducer, but showed varied responses to the second inducer. Those who saw a smaller inducer followed by a larger inducer (i.e., 25 dots labeled as “30” followed by 30 dots labeled as “30”, or 30 dots labeled as “30” followed by 39 dots labeled as “30”) reduced their estimates in the second session, whereas those who saw a larger inducer followed by a smaller inducer were unaffected by the second inducer. Nevertheless, individual differences in the extent of the responses to calibration (translation and scaling parameters of the response grid) were highly correlated between the two sessions across participants (Izard & Dehaene, 2008).

Taken together, although numeral-numerosity mappings are malleable, estimation performance can be resistant to calibration in some individuals and under certain contexts, and highly reliable across conditions in most individuals. This suggests that while an individual's performance across conditions (e.g., uncalibrated versus calibrated, or between two calibrated conditions) might be constrained by similar cognitive mechanisms, there are other cognitive mechanisms which contribute to meaningful individual differences in both the nature and extent of response to calibration.

More recently, Alvarez et al. (2017) provided evidence that 4- and 5-year-old preschoolers' domain-general conceptual analogical skills (e.g., “fish goes with fishbowl just like dog goes with…”) were predictive of the likelihood that a child was affected by calibration or not. They also found that domain-specific numerical analogical skills (e.g., “2 goes with 4 just like 20 goes with…”) predicted how proportionally spaced (i.e., linear) a child's uncalibrated estimates were (Alvarez et al., 2017). Crucially, if affine transformation of a response grid is the mechanism underlying the change between uncalibrated and calibrated verbal estimates, we hypothesize that the transformation of the response grid should manifest as a change in linearity of estimates upon calibration (i.e., estimates may become more proportionally spaced relative to one another). Further, it is plausible that this transformation relies on a combination of prerequisite arithmetic knowledge (e.g., Castronovo & Göbel, 2012), proportional reasoning (e.g., Barth et al., 2016; Barth & Paladino, 2011), and analogical reasoning (e.g., Alvarez et al., 2017; Thompson & Opfer, 2010). Indeed, this hypothesis is consistent with a number of studies that found a relation between math competence and linearity of numerosity estimates in children (Alvarez et al., 2017; Wong, Ho, & Tang, 2016a, 2016b) and adults (Chesney, Bjalkebring, & Peters, 2015). Although most prior work has been correlational, the relations between estimation performance and math competence have typically been interpreted as providing support for a foundational role of stable numerosity-numeral mapping ability in math competence. To the best of our knowledge, no study has examined whether the extent of the calibration effect itself (i.e., changes in estimation performance upon calibration) is associated with math competence, as the response-grid model would suggest, so it is unclear whether the associations between estimation performance and math competence observed were driven by cognitive processes related to calibration (e.g., practice trials with feedback). In summary, math competence may be associated with spontaneous or externally induced local-to-global calibration, possibly more so than numerosity discrimination itself. Such a finding would challenge the relevance of estimation performance measures as predictors of math competence. Moreover, it is unknown whether the initial, uncalibrated estimation performance plays a role in its malleability, and what aspects of performance it might impact. A better understanding of the factors that might facilitate or hinder students' learning and refinement of their estimation skills is critical for designing instruction that can address those factors optimally for each and every student.

3. Current study

The aims of the current study were to test three predictions stemming from the response-grid model by Izard and Dehaene (2008), and the associative mapping – structural mapping model proposed by Sullivan and Barner (2013). Specifically, we set out to investigate (1) the continuity of calibration effects across the numerosities tested at the individual level; (2) the reliability of estimation performance across uncalibrated and calibrated conditions using aggregate performance measures such as accuracy, variability, and linearity of estimates; and (3) to elucidate the factors underlying individual differences in the extent and nature of numerical estimation calibration. To address these questions, we used Sullivan and Barner's (2013) miscalibration paradigm with an “overestimating inducer”. We chose this specific paradigm as it is the only calibration paradigm to our knowledge that has been used to demonstrate the associative and structural mapping mechanisms (Alvarez et al., 2017; Sullivan & Barner, 2013, 2014).

3.1. Hypotheses

3.1.1. Continuity of calibration effects

If the distinction between associative mapping and structural mapping is purely a function of set size and estimates are typically made relative to prior estimates or retrieved directly from a transformed response grid, calibration effects (or the lack thereof) should be reliably continuous across the tested range in most, if not all, participants.

3.1.2. Reliable individual differences in performance across conditions

If estimation performance across calibration conditions is constrained by similar cognitive mechanisms (e.g., discrimination sensitivity, or idiosyncratic but reliable affine transformations of the response grids), we expected that participants' accuracy, variability, and linearity would be correlated across the uncalibrated and calibrated estimation tasks.

3.1.3. Factors underlying participants' responses to calibration

Firstly, the extent to which participants can accurately label quantities with numerals may influence response to calibration. As it is theoretically unclear whether accurate estimators, underestimators, or overestimators during the uncalibrated estimation task may be more or less responsive to the specific calibration used, no prediction seems possible. Secondly, regardless of mapping accuracy, participants with less consistent mappings from trial to trial may be more responsive to calibration as the explicit upper bound may help constrain their estimates. Hence, participants who were initially less consistent in their estimates during the uncalibrated estimation task would shift their estimates to a greater extent (i.e., greater changes in accuracy), or show greater improvements in estimation performance (i.e., reduced variability and/or increased linearity). Thirdly, it is possible that the accuracy and consistency of single numeral-numerosity mappings may not be as critical as the structural coherence and interdependency of mappings across the whole range of numerosities tested (Alvarez et al., 2017). Hence, participants with greater linear structure across their estimates during the uncalibrated estimation task may be more responsive to calibration as they would have demonstrated better spontaneous rescaling of the response grid before calibration. These participants would likely show greater changes in accuracy, reduced variability, or increased linearity.

Moreover, if calibration involves affine transformation of the response grid, math competence may be a potential factor affecting the extent of calibration. Specifically, participants with higher math competence may be better supported in performing the affine transformations of the response grids, and show a greater calibration effect, specifically an increase in linearity of the estimates across the numerosities tested. The response-grid model and the use of a misleading calibration do not allow us to make meaningful predictions of the relation between math competence and the calibration effects on variability and accuracy.

4. Methods

4.1. Participants

Seventy-two undergraduate students (50 female) participated in the study for course credit. The experimental protocol was approved by our Institutional Review Board. All participants provided written informed consent. Data from one participant was excluded due to extreme outlying estimates during the uncalibrated estimation task (see Data Management for further details). Demographic information and the standard scores of standardized math and reading measures of the remaining 71 participants (49 female) are presented in Table 1.

4.2. Procedure

The experiment was conducted during a single session in a quiet room. All participants completed the uncalibrated estimation task, followed by the calibrated estimation task. Each task was self-paced and took approximately 30 min on average. The stimuli for the estimation task were presented using *E*-Prime 2.0 (Psychology Software Tools, Pittsburgh, PA) on a 21.5″ monitor driven at a refresh rate of 60 Hz and a resolution of 1920 × 1080 pixels. The 47.7 × 26.8 cm screen subtended a 43.4° × 25.2° visual angle with an approximate viewing distance of 60 cm. Standardized reading and mathematical tests were administered after the estimation tasks. Finally, a brief questionnaire was administered verbally and informally to ensure that any participant who may have been suspicious about the intentional miscalibration could be identified. Responses were transcribed verbatim.

Table 1
Demographic information and standardized test measures (N = 71).

Measure	Mean	SD	Range
Age (years)	19.63	1.03	18.08–22.17
WCJ-III Calculation	120.77	12.51	89–148
WCJ-III Math Fluency	112.23	13.26	81–151
WCJ-III Reading Fluency	117.82	11.27	89–147

Note. WCJ-III: Woodcock-Johnson III Tests of Achievement.

4.2.1. Uncalibrated estimation task

Our estimation paradigm was adapted from that reported by Sullivan and Barner (2013). Participants saw a randomly ordered series of blue dot arrays presented at the center of a grey circular background (diameter of 23 cm) against a black screen. Each array was 1000 × 1000 pixels, which covered a visual angle of 21.7° × 21.7°, and the diameter of each dot subtended a visual angle of 0.19° (0.2 cm) to 0.48° (0.5 cm). On each trial, the dot arrays were presented for 500 milliseconds (ms), followed by a circular grey mask prompting for an estimate. We opted for a short presentation duration to prevent participants from counting. Previous adult studies have employed presentation times as short as 100 ms (e.g., Izard & Dehaene, 2008) to as long as 1500 ms (e.g., Chesney & Matthews, 2018). Participants were given no information about the range of numerosities they would see and were instructed to estimate the number of dots and record their estimates using the numeric keypad on a computer keyboard as quickly and as accurately as possible. Confirmation of each estimate was made by pressing the spacebar key, upon which a central fixation cross within a circular grey background was then presented for 1500 ms before the next set of dots were presented. Participants were allowed to amend their estimates before confirmation by using the backspace key whenever necessary. Response latencies (measured from the onset of the response screen to the confirmation of their estimates) ranged from 0.29 to 32.2 s (s) (Mean = 2.6 s, SD = 1.6 s). As response latencies were not a pure measure of numerosity encoding, we did not analyze them further. There were no practice trials, and no feedback was given throughout the task.

Fifteen numerosities were presented: 8, 12, 20, 35, 60, 80, 95, 120, 150, 180, 200, 240, 275, 300, and 350 (see Sullivan & Barner, 2013). We deliberately excluded numerosities below 8 to avoid subitizing and minimize counting potential. Each numerosity was presented 18 times, resulting in a total of 270 trials per task. To minimize the use of non-numerical visual cues, each numerosity was matched with every other numerosity on dot size for half the trials, and on total occupied area for the other half (Dehaene, Izard, & Piazza, 2005).

4.2.2. Calibrated estimation task

The stimuli and instructions were identical to those in the uncalibrated estimation task, except that participants were told once, verbally, at the beginning of the task that the largest set they would see was 750. This was chosen based on the findings of Sullivan and Barner (2013): In the uncalibrated condition, the mean estimates were up to about 220. When participants were calibrated to 375 (close to the veridical upper bound of 350), their mean estimates were surprisingly lower (up to about 170). When participants were calibrated to 750, their mean estimates were higher (up to about 260; hence, the "overestimating inducer" of 750 seems appropriate for increasing the accuracy of the estimates in the sampled range. Response latencies ranged from 0.24 to 29.4 s (Mean = 2.2 s, SD = 1.4 s).

4.2.3. Mathematical and reading competencies

Mathematical competence was assessed 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 simple addition, subtraction, and multiplication problems with numerals 0 to 10 as quickly as possible within three minutes. The Calculation subtest is an untimed test including arithmetic (with natural and rational numbers), algebra,

trigonometry, and calculus. While the Math Fluency subtest primarily assesses fluency of arithmetic fact retrieval, Calculation subtest assesses a broader scope of calculation competence comprising conceptual and procedural knowledge. Examining these subtest scores separately allowed us to examine the adequacy of arithmetic fact retrieval in supporting calibration.

To assess the specificity of the relation between numerical estimation performance and mathematical competence, reading competence was assessed using the Reading Fluency subtest of the WCJ-III Tests of Achievement. It requires participants to read a series of sentences as quickly as possible and indicate whether the sentence is true or false within 3 min. Reading Fluency not only serves as a proxy for general cognitive ability, but also a measure of the ability to infer symbol-referent associations fluently. Age-normed standard scores were used for all analyses. Table 1 shows that the sample has a wide and representative range of math and reading scores. All the standardized measures were normally distributed (Shapiro-Wilk; all $ps > .23$). Calculation and Math Fluency subtests scores were moderately correlated [$r(69) = .438$, $p < .001$, $BF_{10} = 185.4$], but neither was correlated with Reading Fluency [$r(69) = .124$, $p = .304$, $BF_{10} = 0.25$, and $r(69) = .191$, $p = .111$, $BF_{10} = 0.52$, respectively].

4.2.4. Manipulation check

A "funnel debriefing" (Bargh & Chartrand, 2000) was administered to assess participants' suspicions of the miscalibration. The procedure began with an abstract, open-ended question about the purpose of the study, followed by more specific questions to probe participants' awareness and suspicion of the miscalibration (e.g., "Did you notice anything unusual about the tasks?"). Participants' responses to the verbally administered survey were transcribed verbatim, and then coded by the first two authors independently. Inter-rater reliability was high (Cohen's $\kappa = .85$). Any discrepancies in coding were resolved through discussion. The questions, coding scheme for response categories, proportions for the response categories, and examples of responses can be found in Supplemental Materials.

4.3. Analyses

4.3.1. Data management

Following previous studies that employed similar paradigms (Alvarez et al., 2017; Sullivan & Barner, 2013, 2014) and to replicate their key group-level findings, we adopted the same criteria for data exclusion. Specifically, we excluded null responses (uncalibrated: $N = 49/19{,}440$ trials; calibrated: $N = 70/19{,}440$), responses of "0" and "1" (uncalibrated: $N = 7/19{,}440$; calibrated: $N = 21/19{,}440$), and responses that were likely to be typing errors, 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: $N = 207/19{,}440$; calibrated: $N = 161/19{,}440$). Within each condition, we further excluded outlying estimates that were more than three standard deviations from the mean of each participant's estimates of each numerosity presented (uncalibrated: $N = 163/19{,}440$; calibrated: $N = 188/19{,}440$). We also visually inspected participants' estimates to assess for any outlying estimates that might have been missed by the trimming procedure described above. For one participant, we excluded two "750" responses in the calibrated task for the second largest numerosity (300). The estimates for that participant did not exceed 200 across both conditions, and those two estimates possibly reflected an occasional need to adhere to the calibrated task instructions, rather than being genuinely representative of the participants' estimates. These two responses also artificially inflated the variability of this participant's estimates for 300, such that its standard deviation was 6.4 times as large as the next largest standard deviation, which was for the largest target numerosity (350). Across the whole sample, 97.81% and 97.74% of the uncalibrated and calibrated data points respectively were retained for further analyses. Even though all participants completed the uncalibrated task followed by the calibrated task, there were no apparent indications that data from the calibrated task contained more errors or outlying data points that may have been attributable to fatigue.

All participants were compliant with the instruction to formulate their estimates based on numerosity as indicated by a significant prediction of their estimates from the target numerosities and typical behavioral signatures expected in numerosity estimation tasks (e.g., scalar variability; see Supplemental Materials). So, no participants were excluded based on this criterion. In the uncalibrated condition, one participant showed an extreme overestimation across the entire range of numerosities (e.g., "3000" in response to 350 dots), resulting in a mean absolute error rate that was more than seven standard deviations from the sample's mean. Hence, we excluded data from this participant from all analyses.

Two other participants had estimates in the calibrated task that were > 750 ($N = 12/19{,}940$ trials), suggesting that participants did not respond as anticipated to the calibration by lowering their estimates within the given range. However, many other participants did not substantially increase their largest estimates to 750 either (see Fig. S1), which also suggested that participants neither associate the largest array seen in the uncalibrated condition with 750, nor uncritically respond with 750 whenever they see a large array. Given the general resistance to what the upper bound truly was (which we will address in our discussion) and that the aim of this study was to examine individual differences in the degree of calibration when presented with the same set of instructions, we did not exclude any other trials or participants based on apparent adoption of, or resistance to, the calibration. Relatedly, we did not find any compelling evidence that participants' suspicions about the misleading calibration would invalidate their data (see Manipulation Check subsection of the Results and Supplemental Materials), hence, we did not exclude participants based on their post-experiment reports.

4.3.2. Estimation metrics

We computed participants' accuracy, variability, and linearity of the estimates for the uncalibrated and calibrated tasks separately (Table 2).

4.3.2.1. Accuracy of estimates. Accuracy was measured by computing the absolute error rate for each data point ($AER = \frac{|\text{Estimate} - \text{Numerosity}|}{\text{Numerosity}}$). We then used the mean AER across all trials as a measure of each participant's overall AER (e.g., Alvarez et al., 2017). Taking the absolute value avoids potential reciprocal cancellation between under- and over-estimation during the computation of its mean. A smaller AER thus reflects greater accuracy, regardless of under- or over-estimation.

Table 2
Descriptive statistics of pre-transformed accuracy, variability, and linearity indices ($N = 71$)

Measure	Mean	Median	SD	Range	Skewness	Kurtosis
Uncalibrated AER	0.46	0.46	0.11	0.26–0.75	0.16	−0.58
Calibrated AER	0.50	0.47	0.24	0.26–1.56	3.08	11.09
Uncalibrated CV	0.31	0.29	0.11	0.17–0.83	2.66	10.23
Calibrated CV	0.32	0.29	0.12	0.15–0.77	1.56	3.62
Uncalibrated R_{lin}^2	0.70	0.70	0.10	0.27–0.83	−1.77	5.33
Calibrated R_{lin}^2	0.68	0.70	0.12	0.28–0.86	−1.29	1.55
AER calibration effect	0.04	−0.002	0.20	−0.18–0.96	3.40	12.83
ER calibration effect	0.19	0.11	0.21	0.02–1.04	2.53	6.88
CV calibration effect	0.01	0.02	0.06	−0.16–0.22	0.24	1.37
R_{lin}^2 calibration effect	−0.02	−0.02	0.08	−0.24–0.16	−0.44	1.13

Note. (A)ER: (Absolute) error rate. CV: Coefficient of variation. Calibration effect = $Estimation\ Index_{Calibrated} - Estimation\ Index_{Uncalibrated}$.

4.3.2.2. Variability of estimates. Variability was measured by computing the coefficient of variation per numerosity ($CV = \frac{\text{Standard deviation of estimates}}{\text{Mean estimate}}$) and taking the mean CV across the range of target numerosities. A smaller CV reflects more consistent estimates on the whole.

4.3.2.3. Linearity of estimates. Linearity (R_{lin}^2) was computed by fitting a simple linear regression model to each participants' trial-level estimates regressed on numerosity as a continuous variable (e.g., Alvarez et al., 2017; Sullivan, Frank, & Barner, 2016). R_{lin}^2 reflects the proportion of the variance in a participant's estimates that can be predicted by accurate knowledge of the number of dots in an array. It thus provides a measure of how the estimates are proportionally spaced relative to one another. Given that all participants gave responses that increased with numerosity within each condition (see Supplemental Materials), a larger R_{lin}^2 reflects better internal, ordinal consistency in their estimates across the numerosity range tested. Although we expected R_{lin}^2 to correlate with CV based on the ordinary least squares approach, it should be noted that it measures more than just variability. For instance, responses with low variability that do not increase with numerosity will yield a small R_{lin}^2.

4.3.3. Measures of calibration effect

Calibration effects were examined for the three indices separately instead of an overall calibration effect, because it is possible that calibration may impact one index (e.g., accuracy), but not another (e.g., variability). We obtained measures of the effect of calibration on each measure using the formula: Calibration effect = $Estimation\ Index_{Calibrated} - Estimation\ Index_{Uncalibrated}$. A positive AER calibration effect indicates an increase in absolute deviation from the target numerosities upon calibration. A positive CV calibration effect indicates an increase in variability. A positive R_{lin}^2 calibration effect indicates an increase in linearity.

As the AERs did not allow us to account for switches between under-estimation and over-estimation (e.g., |-0.2| to |0.2| would result in a null AER calibration effect), and we were interested in the extent participants shifted their estimates in response to the calibration, we computed an additional effect of calibration on accuracy using a different approach. We computed the absolute difference between the mean signed error rates (ER) per numerosity ($|ER_{Calibrated} - ER_{Uncalibrated}|$) to capture switches between any under- and over-estimation, and then computed the mean across all numerosities. For instance, a change in ER from −0.2 to 0.2, or 0.2 to −0.2, will constitute an ER calibration effect of 0.4, indicating a 40% change in the deviation from the target numerosity.

Lastly, as most of these indices were not normally distributed (see Table 2; Shapiro-Wilk; all $ps < .039$), with the exception of the calibration effects on CV (Shapiro-Wilk; $p = .258$) and R_{lin}^2 (Shapiro-Wilk; $p = .051$), we performed a rank-based inverse normal (RIN) transformation on all the non-normal indices. RIN transformation has been found to be the optimal procedure for correlational analyses of non-normal data in terms of controlling for Type I error and improving power over other methods (e.g., nonparametric Spearman's rank-based correlation with untransformed data) (Bishara & Hittner, 2012, 2015). All parametric correlational and regression analyses were conducted on RIN-transformed estimation indices, with the exception of the calibration effects on CV and R_{lin}^2.

4.3.4. Frequentist and Bayesian analyses

Unless otherwise stated, all frequentist analyses presented herein were based on a significance threshold of $p < .05$ (two-tailed). To assess whether the impact of calibration was moderated by numerosity at the group level, and to account for non-independence in the data for the Numerosity (8–350) × Condition (uncalibrated vs. calibrated) repeated-measures design, we used the *nlme* package (Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2017) for R (R Core Team, 2016) to analyze trial-level data using a linear mixed model (LMM). A null (no predictor) by-participant random-intercept model yielded an intraclass correlation coefficient of .262, indicating that about 26% of the total variance in the trial-level estimates could be accounted for by differences among participants, and that a mixed-effects model is appropriate for analyzing the data. Unless otherwise stated, numerosity was treated as a continuous predictor. All planned and post-hoc pairwise comparisons were conducted using the R packages *multcomp* (Hothorn, Bretz, & Westfall, 2008) and *emmeans* (Lenth, 2018).

For individual subject analyses on the interaction between calibration and numerosity, we used both ordinary least squares regression and ANOVA, depending on whether numerosity was modeled as a continuous or categorical variable. For all subject-level fixed-effects models, we assumed unequal variances across numerosities and conditions, and used heteroscedasticity-robust standard errors to make statistical inferences (using R packages *lmtest* (Zeileis & Hothorn, 2002) and *Car* (Fox & Weisberg, 2011)). For all planned pairwise comparisons for each participant (i.e., difference in mean estimates between conditions for each numerosity), the underlying distributions of estimates for each numerosity can be assumed to be normal (Izard & Dehaene, 2008). However, due to the varied trimming of data points for each numerosity per condition, we employed Welch's *t*-tests (Delacre, Lakens, & Leys, 2017; Ruxton, 2006) to account for any unequal number of data points and unequal variances. To account for any potentially non-normal distributions, we also conducted a non-parametric version of the Welch's *t*-test on ranked data (Zimmerman & Zumbo, 1993). This is similar to the Mann-Whitney *U* test, but is less sensitive to unequal variances and sample sizes, although the use of ranked data confers less power when the underlying distributions are indeed normal (Zimmerman & Zumbo, 1993). As we are primarily interested in the continuity of the calibration effects among participants whom we classified as "calibrators", we opted not to correct for multiple comparisons to avoid any apparent discontinuities that may arise from false negatives (i.e., inferring a lack of calibration effect when they exist).

To examine individual differences in participants' responses to calibration, we used aggregate measures as computed above and performed correlational and regression analyses. To control for false positives in our correlational analyses we used Benjamini and Hochberg's (1995) FDR procedure to adjust for multiple comparisons. Uncorrected *p*-values are reported, and whenever applicable, non-significant correlations upon correction are noted. Targeted comparisons of correlation coefficients were analyzed using the R package *cocor* (Diedenhofen & Musch, 2015).

Additionally, to better understand the relative strengths of the relations among the estimation indices as well as math measures, and to provide measurable evidence in support of both positive and null findings, we conducted complementary Bayesian *t*-tests, correlational and regression analyses using JASP 0.8.5 (JASP Team, 2018), jamovi 0.8.1.18 (jamovi project, 2018), and the R package *BayesMed* 1.0.1 (Nuijten, Wetzels, Matzke, Dolan, & Wagenmakers, 2015), and their default "objective" priors (Cauchy distribution scaling factor $r = 0.707$ for *t*-tests, $r = 0.354$ for regression, stretched beta prior width = 1 for correlation). 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 et al., 2017; Wagenmakers 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 > 3, 10, 30, and 100 are considered "moderate", "strong", "very strong", and "extreme" evidence in support of the alternative hypothesis (Jeffreys, 1961; Lee & Wagenmakers, 2013; Wagenmakers, Love, et al., 2017).

5. Results

We first report findings of the manipulation check to assess the extent to which our findings might be influenced by participants' awareness of the miscalibration (see Supplemental Materials for

detailed results). Next, we report the results replicating Sullivan and Barner's (2013) group-level distinction between associative and structural mapping to validate the intended calibration manipulation, and to further characterize individual differences in the associative mapping–structural mapping distinction. Finally, we report the results addressing the reliability of estimation performance across conditions and the predictors of response to calibration. A complete characterization of the performance of each task, demonstrating their validity in eliciting the signature behaviors of numerosity estimation tasks, can be found in the Supplemental Materials.

5.1. Manipulation check

We considered the calibration manipulation a failure for an individual participant if they correctly identified the miscalibration as the purpose of the study. It was unlikely for any participant to know definitively that the largest set presented was 350 dots, and that the calibration was incorrect. Indeed, although 22.1% of the participants mentioned that one of the purposes of the study was to investigate how the calibration instructions would affect their estimates, no participant pointed out that the calibration was intentionally wrong to mislead them (see Supplemental Materials for examples of their responses and further analyses of questions specific to the miscalibration). This suggests that the miscalibration per se was not salient within the context of the whole experimental session. To preview, we did not find strong evidence that any participant was certain about the miscalibration and ignored the calibration as a result. Hence, the calibration manipulation was effective in eliciting the intended effects, and we had no compelling reason to exclude any participant from subsequent analyses.

5.2. Associative versus structural mapping, and their continuity in individuals

5.2.1. Group-level analyses

A linear mixed model (LMM) was fit to predict participants' trial-level estimates from the target numerosity, calibration condition (uncalibrated vs. calibrated) and the interaction between numerosity and calibration as fixed factors, and participant as a random factor. We modeled the maximal random effects structure possible (Barr, Levy, Scheepers, & Tily, 2013) by specifying by-participant random intercepts to account for individual differences in estimation baseline as well as random slopes to account for individual differences in the main effects of numerosity and calibration and their interaction. Across both conditions, participants' estimates increased with the target numerosity [$F(1,37{,}449) = 217.96$, $p < .0001$]. While there was no main effect of calibration [$F(1,37{,}449) = 2.84$, $p = .092$], there was a calibration by numerosity interaction [$F(1,37{,}449) = 22.34$, $p < .001$]. Post-hoc tests revealed an effect of calibration for numerosities 12 through 350 ($ps < .046$), but no effect of calibration for numerosity 8 ($p = .132$; Fig. 2). As there was heteroscedasticity in the residuals (i.e., the residuals increased with numerosity), which might lead to biased standard errors and inferences made from hypothesis testing, we re-analyzed the LMM with log-transformed estimates (i.e., criterion) and numerosity (i.e., predictor) that met the assumption of residual homoscedasticity (e.g., Crollen et al., 2011; Crollen & Seron, 2012; Izard & Dehaene, 2008). A significant, but weaker calibration by numerosity interaction was still observed with log-transformed data, with an effect of calibration even for numerosity 8, albeit of a smaller magnitude relative to that for numerosities 12 and above ($ps < .001$) (see Supplemental Materials, Fig. S3).

To rule out the explanation that calibration could have had a greater effect on large numerosities than on small ones because estimates of large numerosities are less accurate and more variable, we computed a normalized calibration effect measure for each target numerosity ($\frac{Mean\ Estimate_{Calibrated} - Mean\ Estimate_{Uncalibrated}}{Mean\ Estimate_{Uncalibrated}}$), identical to that employed by Sullivan and Barner (2013). This measure captures the proportional rather than the absolute change for each numerosity. To assess the largest numerosity that was unaffected by calibration, we performed a Dunnett's test on the mean normalized calibration effect for each numerosity against a value of zero (see Sullivan & Barner, 2013). Similar to our findings above, there was an effect of calibration for numerosities 12 through 350 ($ps < .001$, Fig. S4), but no effect of calibration for numerosity 8 ($p = .463$). Taken together, these findings were consistent with those of Sullivan and Barner (2013), in that participants appeared to be more influenced by the misleading upper bound for the larger numerosities than for the smaller numerosities.

5.2.2. Individual-level analyses

To characterize the calibration effect at the individual level, we first fit a simple linear regression model to each participant's data to predict the trial-level estimates from the target numerosity, calibration condition, and the calibration by numerosity interaction (see Sullivan & Barner, 2013). Using $p < .05$ as our classification criterion, all participants showed a main effect of numerosity, 48 (67.6%) showed a main effect of calibration, 49 (69.0%) showed a calibration by numerosity interaction, 59 (83.1%) 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); see Fig. S1 for plots of calibrators versus non-calibrators). Of the 59 calibrators, 38 (53.5% of the sample) showed both main effect of calibration and a calibration by numerosity interaction, and the remaining 21 showed either only a main effect of calibration or only an interaction. All proportions reported above were significantly above chance (chance = .05 for each independent main effect, .1 for either main effect or interaction, and .025 for both main effect and interaction), one-sided binomial $p < .001$.

Finally, we examined the smallest numerosity that was affected by the calibration at the individual level among the calibrators, and whether the effects of calibration were continuous up to the largest numerosity presented. Based on the definition of structural mapping, we defined continuity in the calibration effects as pairwise differences between conditions that extend continuously from the smallest numerosity affected by calibration up to numerosity 350. Hence, if a participant showed pairwise differences between conditions for numerosities 240, 300, and 350, but not for 275, we considered that as discontinuous calibration effects. To avoid assuming that the effects of calibration should follow a continuous trend across the numerosities sampled, which is typically assumed in a linear regression framework in previous studies, we modeled numerosity as a categorical variable (e.g., Castronovo & Göbel, 2012; Izard & Dehaene, 2008). To assess the continuity of the calibration effects, for each participant, we conducted a 2 (Condition: Uncalibrated vs. Calibrated) × 15 (Numerosity: 8–350) Analysis of Variance, followed by planned pairwise comparisons between the mean uncalibrated and calibrated estimates for each numerosity using Welch's t-test (dfs = 15.19–34). As a result of modeling numerosity as a categorical predictor instead of a continuous one, four participants switched from being calibrators to non-calibrators, and four vice versa. For the next set of analyses, we focused on the 55 participants who were classified as calibrators regardless of whether numerosity was a categorical or continuous predictor. The smallest numerosity affected by calibration ranged from 8 to 200. Among the calibrators, more of them showed discontinuous calibration effects (82%) than continuous ones (18%) (chance = .5, one-sided binomial $p < .001$). Overall classifications were highly similar with a non-parametric version of the Welch's t-test using ranked estimates (discontinuous calibration effects: 82%, continuous calibration effects: 18%) (dfs = 16–34). Fig. 3 illustrates the nature and extent of discontinuity in calibration effects in four representative calibrators (see Fig. S5 for similar plots for all other participants). To provide additional information about data insensitivity or noise in these analyses, Bayes factors are also plotted in Figs. 3 and S5. BF_{10} approaching 1 is

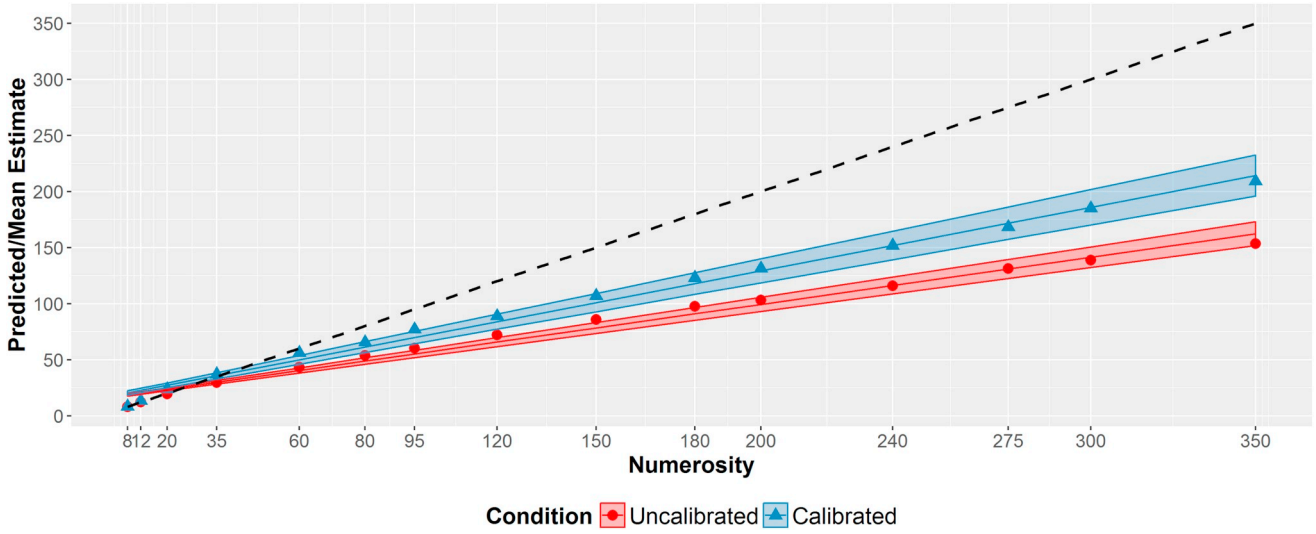

Fig. 2. Predicted estimates (with bands reflecting standard errors of the mean) as a function of numerosity and condition (uncalibrated vs. calibrated) fit by a linear mixed model on trial-level estimates. Data points represent the grand mean as a function of numerosity and condition. Black dashed line represents the veridical estimates.

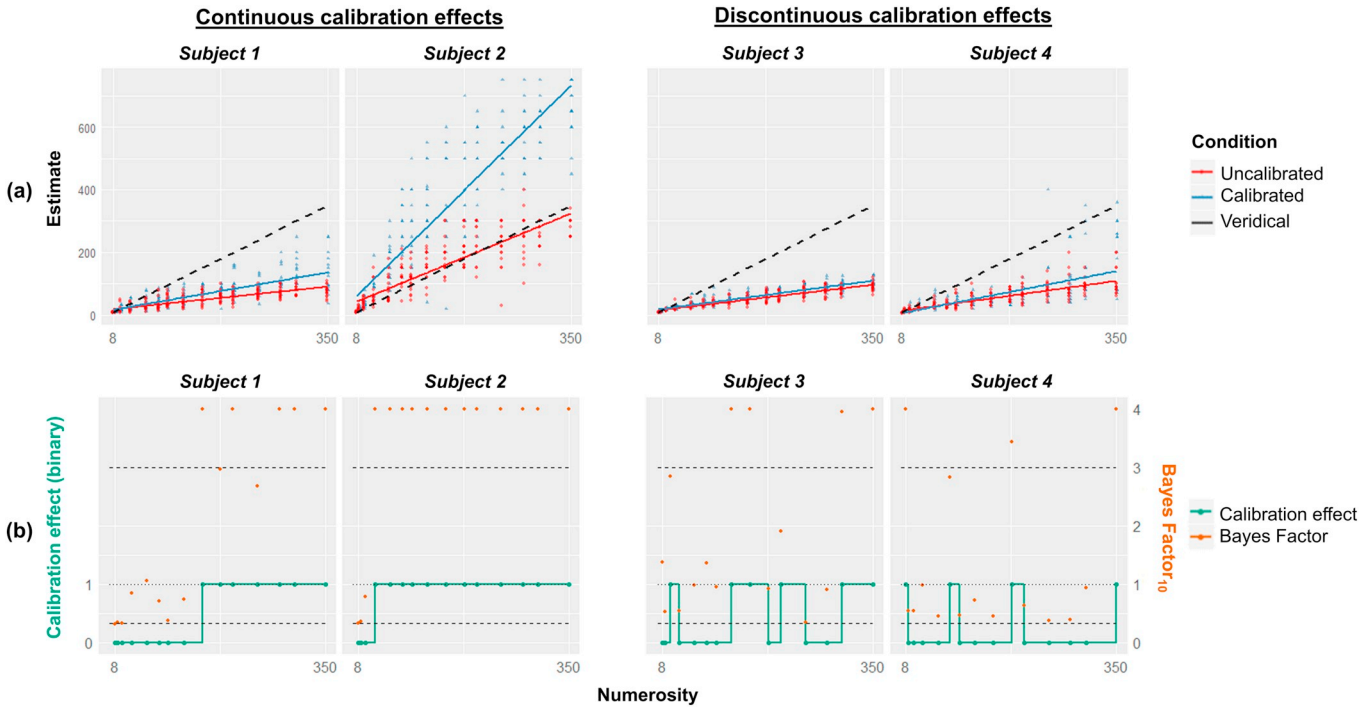

Fig. 3. Examples of continuous and discontinuous calibration effects in four representative participants who showed a calibration by numerosity interaction effect. (a) Raw estimates per condition. (b) Green step-plots reflect binary coding of "1" for significant pairwise difference between conditions per numerosity and "0" for non-significant pairwise difference at $p < .05$, 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 the null and alternative hypotheses. The continuity of the calibration effects for all 71 participants are shown in Fig. S5. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

indicative of data insensitivity rather than conclusive evidence of a calibration effect ($BF_{10} > 3$) or lack thereof ($BF_{10} < 1/3$) (Dienes, 2014). In summary, regardless of the statistical procedure used, our findings demonstrated discontinuity in calibration effects, and that the effects were not consistently and primarily a function of set size.

5.3. Reliability of individual differences in estimation performance

Table 3 shows the zero-order correlations of all the estimation indices, controlled for multiple comparisons. Firstly, as hypothesized, participants' accuracy, variability, and linearity were positively and highly correlated between the uncalibrated and calibrated conditions ($rs > .67$) (Table 3). Secondly, the effects of calibration on accuracy, variability, and linearity were positively correlated with most of the corresponding measures of the calibrated task, but not with the corresponding measures of the uncalibrated task. In other words, a performance index's response to calibration did not seem to depend on that index's uncalibrated state. However, larger absolute changes in

Table 3
Zero-order correlation coefficients of estimation indices ($N = 71$).

Measure	1	2	3	4	5	6	7	8	9	10
1. Uncalibrated AER	–	.013	−.201	**.675*****	−.014	−.266*	−.183	−.216	−.072	−.148
BF_{10}		(0.15)	(0.59)	**(1.10×10^{8})**	(0.15)	(1.75)	(0.46)	(0.74)	(0.18)	(0.31)
2. Uncalibrated CV		–	**−.761*****	.210	**.815*****	**−.592*****	**.313****	**.451*****	.000	.058
BF_{10}			**(5.93×10^{11})**	(0.68)	**(1.10×10^{15})**	**(2.94×10^{5})**	**(4.67)**	**(307.12)**	(0.15)	(0.17)
3. Uncalibrated R_{lin}^2			–	**−.308****	**−.650*****	**.745*****	−.227	−.122	−.068	−.063
BF_{10}				**(4.20)**	**(1.49×10^{7})**	**(9.11×10^{10})**	(0.87)	(0.25)	(0.17)	(0.17)
4. Calibrated AER				–	.079	−.241*	**.521*****	−.063	−.125	.053
BF_{10}					(0.18)	(1.11)	**(6.25×10^{3})**	(0.17)	(0.25)	(0.16)
5. Calibrated CV					–	**−.754*****	.131	**.533*****	**.500*****	**−.336****
BF_{10}						**(2.56×10^{11})**	(0.27)	**(1.10×10^{4})**	**(2.34×10^{3})**	**(8.24)**
6. Calibrated R_{lin}^2						–	−.065	−.120	**−.452*****	**.541*****
BF_{10}							(0.17)	(0.24)	**(321.12)**	**(1.63×10^{4})**
7. AER calibration effect								.165	−.177	.272*
BF_{10}								(0.37)	(0.43)	(1.96)
8. ER calibration effect								–	.224	−.050
BF_{10}									(0.84)	(0.16)
9. CV calibration effect									–	**−.735*****
BF_{10}										**(3.18×10^{10})**
10. R_{lin}^2 calibration effect										–
BF_{10}										

Note. AER: Absolute error rate. CV: Coefficient of variation. Calibration effect = Estimation $Index_{Calibrated}$ − Estimation $Index_{Uncalibrated}$.
$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$. Correlation coefficients in **bold** remained significant after correction for multiple comparisons using Benjamini and Hochberg's (1995) FDR procedure. BF_{10} = Bayes Factor (Alternative/Null hypotheses). Evidence levels: Moderate: $BF_{10} > 3$; Strong: $BF_{10} > 10$; Very strong: $BF_{10} > 30$; Extreme: $BF_{10} > 100$.

accuracy were associated with larger uncalibrated variability, suggesting the possibility of cross-index relations. Lastly, in both uncalibrated and calibrated tasks, as well as the effect of calibration, accuracy was neither associated with variability, nor with linearity, but linearity was negatively correlated with variability. This was expected given that R_{lin}^2 was computed based on the sum of squares of residuals between the observed and predicted data points.

5.4. Predictors of response to calibration

Shown in Table 4, each calibration effect was regressed on the uncalibrated estimation indices, Calculation and Math Fluency, controlling for Reading Fluency. We also report Bayes Factors that provide information of the extent to which the data support the inclusion of a specific predictor of interest, taking into account all possible model combinations with versus without that predictor (Wagenmakers, Love, et al., 2017). Given that variability and linearity were highly correlated, we confirmed that our findings below were not severely affected by collinearity between the variability and linearity indices (variance inflation factors ≤2.63, which were within the commonly used threshold of 10, or even 4; O'Brien, 2007).

5.4.1. Accuracy calibration effect

There were no significant unique predictors for the AER calibration effect. For the ER calibration effect (accounting for cross-sign shifts in ERs upon calibration), the less consistent or more linear participants' uncalibrated estimates were, the more the means of the estimates were shifted upon calibration. However, it should be noted that the bivariate correlation between uncalibrated linearity and ER calibration effect was negative and non-significant ($r = -.122$, $BF_{10} = 0.25$), whereas uncalibrated linearity was a positive and significant predictor ($\beta = .457$, $BF_{10} = 15.18$) in the multiple regression analysis. This change in direction and significance of the contribution of uncalibrated linearity to the ER calibration effect could possibly reflect the unique contribution of linearity upon controlling for the collinearity between variability and linearity. Taken together, uncalibrated variability seemed to play a key and consistent role in the extent of directional changes in accuracy. Linearity may also play a unique role in the overall directional changes in accuracy.

5.4.2. Variability calibration effect

Although Calculation was a significant predictor of a change in variability upon calibration, the full model was not significant. Further analyses showed that the overall model was significant only with Calculation ($\beta = -.272$, $t = -2.35$, $p = .02$, $BF_{10} = 2.48$) as a sole predictor, $F(1,69) = 5.52$, $p = .02$, but not when Reading Fluency and any other predictors were entered. As the F-test of overall significance assesses and controls for multiple coefficients being simultaneously compared, these suggest that the statistical significance of Calculation should be interpreted with caution.

5.4.3. Linearity calibration effect

Participants with higher Calculation scores showed greater increase in linearity upon calibration. Math Fluency, however, did not predict the response of linearity to calibration. To test the specificity of the relation between Calculation and the change in linearity, we ran a separate model with all other predictors except Calculation in the first step. Calculation was still uniquely predictive of the change in linearity ($\beta = .422$, $p = .001$, $BF_{10} = 27.47$) over and above all other predictors in the null model ($\Delta R^2 = .137$, $p = .001$). As the calibration effects on variability and linearity were highly correlated, we further controlled for CV calibration effect in the null model. Calculation was still uniquely predictive of the change in linearity ($\beta = .212$, $p = .015$, $BF_{10} = 3.68$) even after further controlling for CV calibration effect ($\Delta R^2 = .032$, $p = .015$).

5.4.4. Math competence as predictors of linearity of uncalibrated and calibrated estimates

Finally, we asked whether Calculation scores were associated not only with the change in linearity, but also with the initial (uncalibrated) and final (calibrated) linearity of the estimates. This addressed whether associations between linearity of estimates and math competence observed in previous studies might have been driven by cognitive processes related to some explicit or implicit calibration. To this end, we computed partial correlations between Calculation scores and the linearity indices for the uncalibrated and calibrated conditions, controlling for Reading Fluency scores. Calculation scores correlated positively with linearity of the calibrated estimates [$r(68) = .282$, $p = .018$, $BF_{10} = 2.31$], but not with linearity of uncalibrated estimates [$r(68) = .129$, $p = .287$, $BF_{10} = 0.25$]. These correlations were

Table 4
Hierarchical regression analyses predicting the calibration effects on accuracy, variability, and linearity from uncalibrated estimation indices, Math Fluency and Calculation, controlling for Reading Fluency ($N = 71$).

Dependent measure/Step	Predictors	β	R^2	ΔR^2	$BF_{10\ (Inclusion)}$
AER calibration effect					
Step 1	Reading Fluency	.083	.007		0.30
Step 2			.162	.155*	1.17
	Reading Fluency	.094			
	Uncalibrated AER	−.187			1.30
	Uncalibrated CV	.280			3.13
	Uncalibrated R_{lin}^2	−.073			0.69
	Math Fluency	−.021			0.42
	Calculation	.138			0.70
ER calibration effect					
Step 1	Reading Fluency	.059	.003		0.27
Step 2			.345***	.341***	842.92
	Reading Fluency	.045			
	Uncalibrated AER	−.129			0.69
	Uncalibrated CV	.808***			3231.54
	Uncalibrated R_{lin}^2	.457**			15.18
	Math Fluency	.057			0.39
	Calculation	.033			0.38
CV calibration effect					
Step 1	Reading Fluency	.036	.001		0.25
Step 2			.121	.120	0.43
	Reading Fluency	.107			
	Uncalibrated AER	−.179			0.67
	Uncalibrated CV	−.217			0.49
	Uncalibrated R_{lin}^2	−.233			0.51
	Math Fluency	−.061			0.55
	Calculation	−.290*			3.20
R_{lin}^2 calibration effect					
Step 1	Reading Fluency	.090	.008		0.31
Step 2			.214*	.206**	5.81
	Reading Fluency	.051			
	Uncalibrated AER	−.102			0.47
	Uncalibrated CV	.031			0.51
	Uncalibrated R_{lin}^2	−.119			0.55
	Math Fluency	.013			0.37
	Calculation	.422**			71.09

Note. (A)ER: (Absolute) error rate. CV: Coefficient of variation.
* $p < .05$. ** $p < .01$. *** $p < .001$.
The Bayes Factors reported for the individual predictors were averaged across all possible candidate models that included the specific predictor of interest versus all other models that omitted the predictor of interest (Wagenmakers, Love, et al., 2017). The null Bayesian model in Step 2 includes Reading Fluency. BF10 = Bayes Factor (Alternative/Null hypotheses). Evidence levels: Moderate: BF10 > 3; Strong: BF10 > 10; Very strong: BF10 > 30; Extreme: BF10 > 100.

statistically different based on a suite of nine statistical comparisons (*cocor* in *R*, Diedenhofen & Musch, 2015), all *p*s ≤ .0352 (one-sided) with the exception of 95% confidence interval tests by Zou (2007) [−.011, .315] and Meng, Rosenthal, and Rubin (1992) [−.013, .334]. To assess the specificity of Calculation, we performed identical analyses with Math Fluency. Math Fluency scores were not associated with either initial or final linearity of the estimates [$r(68) = .031$, $p = .801$, $BF_{10} = 0.15$; $r(68) = .144$, $p = .234$, $BF_{10} = 0.29$, respectively]. In sum, Calculation scores specifically were associated with the change in and final linearity of the estimates, but not with initial linearity.

6. Discussion

It is common to incorporate calibration in instruction to initialize or improve estimation skills in research (e.g., Barth et al., 2016; Kucian et al., 2011; Opfer & Siegler, 2007; Opfer & Thompson, 2008; Peeters, Sekeris, Verschaffel, & Luwel, 2017; Peeters, Verschaffel, & Luwel, 2017; Piazza, Pinel, Le Bihan, & Dehaene, 2007; Revkin et al., 2008; Thompson & Opfer, 2008, 2010) and classroom settings (e.g., Crites, 1993; Joram, Subrahmanyam, & Gelman, 1998; Siegel, Goldsmith, & Madson, 1982; Van de Walle & Thompson, 1985). However, individuals vary greatly in their responses to calibration during numerosity estimation tasks, and little is known about the mechanisms underlying the nature and extent of calibration across individuals. The current study examined these individual differences and explored their underlying factors, with a focus on the roles of estimation-specific factors (i.e., uncalibrated accuracy, variability, and linearity) and math competence. This study thus provides the first step in understanding why some students may not be able to acquire and apply estimation strategies as optimally and as quickly as others.

6.1. Individual differences in calibration continuity

Although the distinction of mapping types as a function of set size was clearly observed at the group level, replicating previous studies (Alvarez et al., 2017; Sullivan & Barner, 2013, 2014), individual-level analyses revealed that it was not set size dependent. In particular, the effects of calibration did not extend continuously beyond the smallest affected numerosity across most participants, contrary to the notions of "global calibration" and "structural mapping", as well as the response-grid model. In fact, discontinuity appeared to be the more common phenomenon. Such discontinuities could be due to relatively strong associative mappings for large numbers, which can be observed in the natural environment (Dehaene & Mehler, 1992), as well as in studies that trained participants to associate large numerosities with novel shapes (Lyons & Ansari, 2009; Lyons & Beilock, 2009; Malone, Heron-delaney, Burgoyne, & Hulme, 2019; Merkley & Scerif, 2015; Merkley, Shimi, & Scerif, 2016; Zhao et al., 2012). Alternatively, discontinuities may arise from trial-to-trial strategy variation (e.g., Crites, 1992; Gandini, Ardiale, & Lemaire, 2010; Gandini, Lemaire, & Dufau, 2008; Luwel, Lemaire, & Verschaffel, 2005). Although it is possible that the observed discontinuities may be due to noisy data at the numerosity level, Bayesian analyses provided evidence that noise does not fully account for all discontinuities. The existence of noise amidst strong calibration effects suffices to call into question the continuity assumption of "structural mapping" as well as the response-grid model.

Hence, these findings suggest a reconsideration of the very concept of stable and direct mappings between symbolic and nonsymbolic numerosity representations (Dehaene, 2007; Piazza, 2010; Piazza & Eger, 2016; Stoianov, 2014). In fact, the response-grid model already suggests that the mappings are malleable in the presence of an external calibration. Perhaps the spontaneous response grid is relatively stable, but to our knowledge, test-retest *stability* of uncalibrated estimation performance has not be empirically demonstrated. In any case, the model falls short of accounting for the discontinuity in calibration effects, and does not take into account the possibility of strategic variation as observed in numerous studies. Other cognitive strategies may play a role in supporting on-the-fly item-by-item as opposed to system-level mappings. For instance, Chesney and Matthews (2018) propose an item-level "relational" mechanism in that a perceptual sense of proportion between *two* sets (e.g., if 25 dots ≈ "20", then 50 dots ≈ "40") may facilitate more accurate numeral assignment for unfamiliar large sets than relying on a direct mapping between *one* set and its corresponding numeral (see also Alvarez et al., 2017). This is supported by the highly linear and accurate performances on a ratio estimation task (estimate the ratio instantiated by a pair of dot sets) and a nonsymbolic version of a number-line estimation task (estimate the relative position of a dot set on a line bounded by two dot sets as anchors), relative to a severe underestimation in an uncalibrated numerosity estimation task (Chesney & Matthews, 2018).

6.2. Reliability of individual differences in performance

Participants' accuracy, variability, and linearity of their estimates were highly correlated across conditions. This suggests that similar

cognitive mechanisms may be driving and constraining performance on both the uncalibrated and calibrated estimation tasks. One candidate mechanism could be the perceptual encoding and discrimination of numerosities (i.e., encoding stage in Fig. 1). In having participants complete the same estimation task with two different calibration inducers (e.g., 25 dots labeled as "30" followed by 30 dots labeled as "30"), Izard and Dehaene (2008) observed that the internal Weber fraction – a measure of the amount of noise inherent to the mental representations of numerosities – remained stable across two calibrated conditions (R^2 = .61) and was unaffected by the nature of the calibration inducers. By definition, the Weber fraction should be comparable to the coefficient of variation (CV) as they are both measures of the noise of a given numerosity representation that is proportional to the numerosity itself (Chesney et al., 2015; Guillaume, Gevers, & Content, 2016). Indeed, we observed a strong correlation (r = .8) between the CVs of the uncalibrated and calibrated conditions. Although CV is typically correlated with the Weber fraction (Castronovo & Göbel, 2012; Libertus, Odic, Feigenson, & Halberda, 2016; Pinheiro-Chagas et al., 2014; Wong et al., 2016b), CV from any estimation task involving numerals is not a pure measure of the acuity of the mental number line (Ebersbach, Luwel, & Verschaffel, 2013) as it also encompasses the response biases (i.e., shifting and scaling) in symbolic labeling, and may not even correlate with Weber fraction in some instances (Guillaume et al., 2016).

If a response grid exists, another mechanism driving estimation performance may involve the affine transformations supporting the generation of *calibrated* response grids from the *spontaneous* response grid (i.e., response selection stage in Fig. 1). Izard and Dehaene (2008) found that regardless of whether the calibration was spontaneous or externally induced, participants tended to stretch or compress, and translate to similar relative extents. In the current study, it is possible that the influence of both of these mechanisms manifested in reliable differences in estimation performance across conditions.

6.3. Roles of uncalibrated variability and linearity of estimates in changes in accuracy

We found that participants with less variable numerosity-to-numeral mappings were less likely to shift the distributions of their estimates. A possible explanation for this behavior is that the lower variability could reflect stronger "associative mappings" and direct memory retrieval across a wide range of numerosities, possibly due to more distinctive mental representations of the numerosities. These stronger associative mappings would therefore be less susceptible to calibration. An alternative explanation is that the variability of participants' numeral-numerosity mappings may reflect their confidence in their overall estimation abilities (Halberda & Odic, 2015; Libertus et al., 2016), and greater confidence would lead to lower susceptibility to external calibration. It would be informative for future studies to examine these non-mutually exclusive hypotheses, possibly using measures of trial-level strategies and confidence ratings.

Additionally, we found that participants with higher uncalibrated linearity (i.e., more proportionally spaced the estimates were relative to one another) shifted the distribution of their estimates to a greater extent. A possible explanation for this is that participants with a more internally coherent structure could modify their estimates *systematically across the entire range* as compared to participants with a less internally coherent structure, leading to a greater effect of calibration across the entire range of numerosities. Taken together, the extent to which participants responded to the calibration by shifting the distributions of their estimates may depend on how variable or linear their uncalibrated estimates were.

6.4. Role of calculation competence in changes in linearity

Consistent with both the response grid model (Izard & Dehaene, 2008) and an analogy-based "structural mapping" (e.g., ::: is to :::::: as 6 is to…?) (Alvarez et al., 2017), we found that participants with higher calculation competence showed a greater increase in the linearity of their estimates. Notably, fluency of arithmetic fact retrieval did not seem as critical as broad calculation competence. Moreover, although individual differences in linearity of estimates were highly reliable across the uncalibrated and calibrated conditions, suggesting that they may tap into the same underlying representations, their relation to calculation competence differed significantly. It is therefore possible that the associations found between estimation performance and math competence observed in previous studies might have been partly driven by participants' ability to calibrate their estimates spontaneously or with an external calibration, rather than participants' representations of numerosities per se. Considering that calibration effects may not be reliably continuous at the individual level, it is possible that the role of calculation competence may not be targeted at the system level (i.e., entire response grid), but at a more regional level (e.g., discontinuous segments of a response grid) or trial level.

More broadly, the calibration-specific association between changes in linearity and calculation competence is also consistent with the recent debate on whether different cognitive constructs are measured by bounded (e.g., 0–1000) and unbounded (e.g., 0–?) number-line estimation tasks (in which participants are typically asked to mark a position on a line based on a given numeral, or to assign a numeral given a position marked on a line) (Chesney & Matthews, 2018; Cohen & Blanc-Goldhammer, 2011; Cohen & Sarnecka, 2014; Ebersbach et al., 2013; Ebersbach, Luwel, & Verschaffel, 2015; Kim & Opfer, 2017; Link, Huber, Nuerk, & Moeller, 2014; Link, Nuerk, & Moeller, 2014; Reinert, Huber, Nuerk, & Moeller, 2015; see Schneider et al., 2018, for a meta-analysis). In particular, Cohen and Sarnecka (2014) found age-related changes in children's performance on a bounded number-line task, but not on an unbounded version. Their findings suggest that the changes in bounded-estimation performance may reflect the growth of task-specific measurement skills rather than changes in representations of numerosity (Cohen & Sarnecka, 2014). Several studies (Chesney & Matthews, 2013; Huber, Moeller, & Nuerk, 2014) also found that adults can be easily manipulated to produce estimates on a number-line task that readily fit various linear and non-linear functions, suggesting that number-line estimation may measure transient rather than stable mental representations of numerosity. The current findings also support the notion of transient mental representations of numerosity being measured in the presence of an external calibration and provide evidence that the calibration process may be supported by other cognitive factors such as calculation competence.

In sum, our findings suggest that math competence supports estimation ability rather than, or in addition to, estimation ability (and the underlying mental representations of numerosity) being foundational for math competence (see also Castronovo & Göbel, 2012). Researchers should be aware of prevailing concerns regarding the constructs that uncalibrated and calibrated tasks actually measure. Particularly, our findings highlight that interpreting estimation performance solely as a measure of mental representations of numerosity, without considering strategies and other factors influencing estimation behaviors, may be too restrictive.

7. Limitations

One limitation of the current study is the validity and generalizability of the use of a misleading calibration. In the current study, we are particularly interested in understanding the proposed associative and structural mechanisms underlying the numerosity-to-numeral mappings and to induce calibration in as many participants as possible to examine individual differences. It was therefore critical to find an ideal inducer value that is sufficiently deviant to dissociate mappings that are resistant to calibration from those that are not. Although participants tend to underestimate large numerosities, many are not aware

of such underestimating tendencies and the severity of their own underestimation; yet, they readily calibrate their estimates when an external calibration is available, regardless of its accuracy. For instance, Izard and Dehaene (2008) reported that the calibrated participants "consciously corrected their responses to match the inducer, but when [the experimenters] told [the calibrated participants] that non-calibrated participants had estimated the maximum numerosity at 50 instead of 100, they did not admit that their spontaneous responses would have been that inaccurate" (p. 1234). Moreover, the extent of underestimation, confidence in one's estimates, and awareness of such underestimating tendencies after being told the upper bound are largely idiosyncratic. Hence, while one calibration inducer value would work for one participant, it may not work for another. The inducer used here deviated from the actual numerosity by a factor of about 2 (750 vs. 350), but it is not uncommon for adults' uncalibrated estimates to deviate by as large as a factor of 4 (Minturn & Reese, 1951). In the current study (excluding the outlying participant), the deviation of participants' uncalibrated estimates for numerosity 350 ranged from a factor of 1 to 6.4 (mean = 2.1, median = 1.75). Moreover, using an accurate inducer value does not necessarily guarantee the intended calibration (see Supplemental Materials for additional discussion of findings by Izard & Dehaene, 2008, and Sullivan & Barner, 2013). Although 22% of the participants mentioned that the purpose of the study was related to how well they responded to the calibration instruction, none of them specifically noted that the calibration was intentionally wrong. Analyses of the post-experiment questionnaire responses revealed that some participants acknowledged that they "realized" the miscalibration at some point during the experiment, but their acknowledgements tended to be due to the leading nature of the questions or to hindsight bias (see Supplemental Materials).

Previous studies using the same paradigm had success rates of eliciting calibration effects ranging from 47 to 60% in children (Alvarez et al., 2017; Sullivan & Barner, 2014) to 70–90% in adults (Sullivan & Barner, 2013), and the current study has a success rate of 83%. Taken together with our manipulation checks, this indicates that the paradigm is effective, at least in adults. Even though calibration was not induced in 17% of our participants, they still provide valuable data for analyses of individual differences, which is a broad aim of the current study. In particular, our findings provide preliminary insights into the factors that might underlie a lack of calibration effect on various aspects of estimation performance.

Related to the concern above, because participants' response to calibration likely depends in part on the calibration inducer value relative to the range of numerosities tested, the calibrated estimation task was likely measuring a *state* rather than the *trait* of a participant. Yet, as the performance on the uncalibrated and calibrated tasks were highly correlated, the calibrated task possibly involved both trait and state influences, which Izard and Dehaene (2008) have previously shown. We believe that the present findings are still informative regarding individual differences in the malleability of estimation performance. Future research should use multiple calibration inducer values to examine the intra-subject reliability, and malleability or stability of estimation performance.

Finally, it is possible that calibration effects might not reflect calibration per se, but could reflect practice- or learning-related effects. The current study, however, lacked a no-calibration control group to rule this out. The main aims of this study were not to establish a group-level calibration effect (which was for replication of previous findings), but to examine the effects at the individual level. A control group was therefore not critical for our individual differences analyses. Nonetheless, if there were such practice effects, estimates should become more accurate over time within each condition rather than show an abrupt change in accuracy upon calibration. However, the prevalence of such practice effects was very low (see Supplemental Materials). Hence, the changes in performance between conditions were likely to be abruptly induced by the calibration rather than gradually induced by practice or learning. The relative small influence of practice effects has also been observed in tasks that tap into very similar mechanisms (e.g., numerosity comparison tasks such as in DeWind & Brannon, 2012). Taken together, our study has high success of calibration manipulation, but we acknowledge the limitations and challenges of existing estimation paradigms involving calibration.

8. Conclusions

The current study explored the factors underlying individual differences in the extent and nature of the malleability of numerosity estimation performance. By having participants complete both uncalibrated and calibrated estimation tasks, we observed large, but reliable individual differences in performance across conditions, suggesting that an individual's estimation performance might be consistently constrained by cognitive mechanisms shared across calibration conditions. Contrary to previous findings, discontinuous calibration effects across a range of numerosities were more commonly observed than continuous calibration effects, suggesting that a systemic calibration is more nuanced than previously thought. We also found that the more variable or proportionally spaced (i.e., more linear) participants' uncalibrated estimates were, the greater they shifted the distributions of their estimates upon calibration. Importantly, higher calculation competence, but not fluency in arithmetic fact retrieval, was uniquely associated with an increase in linearity of participants' estimates upon calibration. This finding, and the discontinuity in calibration effects, support the growing evidence that the mappings between numerical symbols and nonsymbolic numerosity representations may not be stable and direct, but transient and mediated by other mechanisms. Moreover, numerosity estimation tasks should not be used as a pure measure of number representations. Taken together, both estimation-specific factors and calculation competence may underlie individuals' responses to calibration, which provide us with insights into individual differences in the relation between estimation and calculation skills.

Funding

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

Declaration of Competing 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.

Acknowledgements

We would like to thank Jessica Sullivan for her advice on methodology and linear mixed model analyses. We would like also to thank Rachel Christ, Yuki (Chengxin) Hu, Jordann Lewis, Jordan Barone, Gideon Ticho, Zachary Berkowitz, and Mary Liz Kim 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).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.actpsy.2019.102877.

References

Alvarez, J., Abdul-Chani, M., Deutchman, P., DiBiasie, K., Iannucci, J., Lipstein, R., ... Sullivan, J. (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.
Bargh, J. A., & Chartrand, T. L. (2000). The mind in the middle: A practical guide to priming and automaticity research. *Handbook of research methods in social and personality psychology* (pp. 253–285). . https://doi.org/10.1163/_afco_asc_1327.
Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001.
Barth, H., & Paladino, A. M. (2011). The development of numerical estimation: Evidence against a representational shift. *Developmental Science, 14*(1), 125–135. https://doi.org/10.1111/j.1467-7687.2010.00962.x.
Barth, H., Slusser, E., Kanjlia, S., Garcia, J., Taggart, J., & Chase, E. (2016). How feedback improves children's numerical estimation. *Psychonomic Bulletin & Review, 23*(4), 1198–1205. https://doi.org/10.3758/s13423-015-0984-3.
Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. http://doi.org/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.
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.
Chesney, D. L., & Matthews, P. G. (2013). Knowledge on the line: Manipulating beliefs about the magnitudes of symbolic numbers affects the linearity of line estimation tasks. *Psychonomic Bulletin & Review, 20*(6), 1146–1153. https://doi.org/10.3758/s13423-013-0446-8.
Chesney, D. L., & Matthews, P. G. (2018). Task constraints affect mapping from approximate number system estimates to symbolic numbers. *Frontiers in Psychology, 9*(October), 1–11. https://doi.org/10.3389/fpsyg.2018.01801.
Cohen, D. J., & Blanc-Goldhammer, D. (2011). Numerical bias in bounded and unbounded number line tasks. *Psychonomic Bulletin & Review, 18*(2), 331–338. https://doi.org/10.3758/s13423-011-0059-z.
Cohen, D. J., & Sarnecka, B. W. (2014). Children's number-line estimation shows development of measurement skills (not number representations). *Developmental Psychology, 50*(6), 1640–1652. https://doi.org/10.1037/a0035901.
Cordes, S., Gelman, R., Gallistel, C. R., & Whalen, J. (2001). Variability signatures distinguish verbal from nonverbal counting for both large and small numbers. *Psychonomic Bulletin & Review, 8*(4), 698–707. https://doi.org/10.3758/BF03196206.
Crites, T. W. (1992). Skilled and less skilled estimators' strategies for estimating discrete quantities. *The Elementary School Journal, 92*(5), 601–619. Retrieved from http://www.jstor.org/stable/1001741.
Crites, T. W. (1993). Strategies for estimating discrete quantities. *Arithmetic Teacher, 41*(2), 106–108.
Crollen, V., Castronovo, J., & Seron, X. (2011). Under- and over-estimation: A bi-directional mapping process between symbolic and non-symbolic representations of number? *Experimental Psychology, 58*(1), 39–49. https://doi.org/10.1027/1618-3169/a000064.
Crollen, V., & Seron, X. (2012). Over-estimation in numerosity estimation tasks: More than an attentional bias? *Acta Psychologica, 140*(3), 246–251. https://doi.org/10.1016/j.actpsy.2012.05.003.
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: Har. http://doi.org/https://doi.org/10.1093/acprof:oso/9780199231447.003.0024
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.
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.
DeWind, N. K., & Brannon, E. M. (2012). Malleability of the approximate number system: Effects of feedback and training. *Frontiers in Human Neuroscience, 6*(68), 1–10. https://doi.org/10.3389/fnhum.2012.00068.
Diedenhofen, B., & Musch, J. (2015). Cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS One, 10*(4), e0121945. https://doi.org/10.1371/journal.pone.0121945.
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., Luwel, K., & Verschaffel, L. (2013). Comparing apples and pears in studies on magnitude estimations. *Frontiers in Psychology, 4*(June), 332. https://doi.org/10.3389/fpsyg.2013.00332.
Ebersbach, M., Luwel, K., & Verschaffel, L. (2015). The relationship between children's familiarity with numbers and their performance in bounded and unbounded number line estimations. *Mathematical Thinking and Learning, 17*(2–3), 136–154. https://doi.org/10.1080/10986065.2015.1016813.
Fox, J., & Weisberg, S. (2011). *An {R} companion to applied regression (second)*. Thousand Oaks {CA}: Sage. Retrieved from http://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 http://cpl.revues.org/4990.
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.
Guillaume, M., Gevers, W., & Content, A. (2016). Assessing the approximate number system: No relation between numerical comparison and estimation tasks. *Psychological Research, 80*(2), 248–258. https://doi.org/10.1007/s00426-015-0657-x.
Halberda, J., & Odic, D. (2015). The precision and internal confidence of our approximate number thoughts. *Evolutionary origins and early development of number processing. Vol. 1. Evolutionary origins and early development of number processing* (pp. 305–333). . https://doi.org/10.1016/B978-0-12-420133-0.00012-0.
Hogan, T. P., & Brezinski, K. L. (2003). Quantitative estimation: One, two, or three abilities? *Mathematical Thinking and Learning, 5*(4), 259–280. https://doi.org/10.1207/S15327833MTL0504_02.
Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal, 50*(3), 346–363.
Huber, S., Moeller, K., & Nuerk, H.-C. (2014). Dissociating number line estimations from underlying numerical representations. *The Quarterly Journal of Experimental Psychology, 67*(5), 991–1003. https://doi.org/10.1080/17470218.2013.838974.
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.
jamovi project (2018). Jamovi (version 0.9)[computer software]. Retrieved from https://www.jamovi.org/.
JASP Team (2018). JASP (version 0.9.0.1)[computer software]. Retrieved from https://jasp-stats.org/.
Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.
Joram, E., Subrahmanyam, K., & Gelman, R. (1998). Measurement estimation: Learning to map the route from number to quantity and Back. *Review of Educational Research, 68*(4), 413–449. https://doi.org/10.3102/00346543068004413.
Kim, D., & Opfer, J. E. (2017). A unified framework for bounded and unbounded numerical estimation. *Developmental Psychology, 53*(6), 1088–1097. https://doi.org/10.1037/dev0000305.
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.
Kucian, K., Grond, U., Rotzer, S., Henzi, B., Schönmann, C., Plangger, F., ... von Aster, M. (2011). Mental number line training in children with developmental dyscalculia. *NeuroImage, 57*(3), 782–795. https://doi.org/10.1016/j.neuroimage.2011.01.070.
Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling. Bayesian cognitive modeling: A practical course.* Cambridge: Cambridge University Presshttps://doi.org/10.1017/CBO9781139087759.
Lenth, R. (2018). Emmeans: Estimated marginal means, aka least-squares means. Retrieved from https://cran.r-project.org/package=emmeans.
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.
Link, T., Huber, S., Nuerk, H.-C., & Moeller, K. (2014). Unbounding the mental number line-new evidence on children's spatial representation of numbers. *Frontiers in Psychology, 4*(JAN), 1–12. https://doi.org/10.3389/fpsyg.2013.01021.
Link, T., Nuerk, H.-C., & Moeller, K. (2014). On the relation between the mental number line and arithmetic competencies. *The Quarterly Journal of Experimental Psychology, 67*(8), 1597–1613. https://doi.org/10.1080/17470218.2014.892517.
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.
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.
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.
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 http://doi.org/S0010027719300320.
Meng, X., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin, 111*(1), 172–175. https://doi.org/10.1037/0033-2909.111.1.172.
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.
Nuijten, M. B., Wetzels, R., Matzke, D., Dolan, C. V., & Wagenmakers, E.-J. (2015). BayesMed: Default Bayesian hypothesis tests for correlation, partial correlation, and mediation. (version 1.0.1)[computer software]. Retrieved from http://cran.r-project.org/package=BayesMed.
O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality and Quantity, 41*(5), 673–690. https://doi.org/10.1007/s11135-006-9018-6.
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.
Opfer, J. E., & Thompson, C. A. (2008). The trouble with transfer: Insights from microgenetic changes in the representation of numerical magnitude. *Child Development, 79*(3), 788–804. https://doi.org/10.1111/j.1467-8624.2008.01158.x.
Peeters, D., Sekeris, E., Verschaffel, L., & Luwel, K. (2017). Evaluating the effect of labeled benchmarks on children's number line estimation performance and strategy use. *Frontiers in Psychology, 8*(June), 1–10. https://doi.org/10.3389/fpsyg.2017.01082.
Peeters, D., Verschaffel, L., & Luwel, K. (2017). Benchmark-based strategies in whole number line estimation. *British Journal of Psychology,* 1–19. https://doi.org/10.1111/bjop.12233.
Piazza, M. (2010). Neurocognitive start-up tools for symbolic number representations. *Trends in Cognitive Sciences, 14*(12), 542–551. https://doi.org/10.1016/j.tics.2010.09.008.
Piazza, M., & Eger, E. (2016). Neural foundations and functional specificity of number representations. *Neuropsychologia, 83*, 257–273. https://doi.org/10.1016/j.neuropsychologia.2015.09.025.
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.
Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team (2017). Nlme: Linear and nonlinear mixed effects models. R package version 3.1–131. Retrieved from https://cran.r-project.org/package=nlme.
Pinheiro-Chagas, P., Wood, G., Knops, A., Krinzinger, H., Lonnemann, J., Starling-Alves, I., … Haase, V. G. (2014). In how many ways is the approximate number system associated with exact calculation? PLoS One, 9(11), e111155. http://doi.org/https://doi.org/10.1371/journal.pone.0111155
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.
R Core Team (2016). R: A language and environment for statistical computing. Vienna, Austria. Retrieved from https://www.r-project.org/.
Reinert, R. M., Huber, S., Nuerk, H.-C., & Moeller, K. (2015). Multiplication facts and the mental number line: Evidence from unbounded number line estimation. *Psychological Research, 79*(1), 95–103. https://doi.org/10.1007/s00426-013-0538-0.
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.
Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test. *Behavioral Ecology, 17*(4), 688–690. https://doi.org/10.1093/beheco/ark016.
Schneider, M., Merz, S., Stricker, J., De Smedt, B., Torbeyns, J., Verschaffel, L., & Luwel, K. (2018). Associations of number line estimation with mathematical competence: A meta-analysis. *Child Development, 89*(5), 1467–1484. https://doi.org/10.1111/cdev.13068.
Siegel, A. W., Goldsmith, L. T., & Madson, C. R. (1982). Skill in estimation problems of extent and numerosity. *Journal for Research in Mathematics Education, 13*(3), 211–232.
Stevens, S. S. (1957). On the psychophysical law. *Psychological Review, 64*(3), 153–181. https://doi.org/10.1037/h0046162.
Stoianov, I. (2014). Generative processing underlies the mutual enhancement of arithmetic fluency and math-grounding number sense. Frontiers in Psychology, 5(NOV), 1–4. http://doi.org/https://doi.org/10.3389/fpsyg.2014.01326
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.
Thompson, C. A., & Opfer, J. E. (2008). Costs and benefits of representational change: Effects of context on age and sex differences in symbolic magnitude estimation. *Journal of Experimental Child Psychology, 101*(1), 20–51. https://doi.org/10.1016/j.jecp.2008.02.003.
Thompson, C. A., & Opfer, J. E. (2010). How 15 hundred is like 15 cherries: Effect of progressive alignment on representational changes in numerical cognition. *Child Development, 81*(6), 1768–1786. https://doi.org/10.1111/j.1467-8624.2010.01509.x.
Van de Walle, J., & Thompson, C. S. (1985). Estimate how much. *Arithmetic Teacher, 32*(9), 4–8. Retrieved from http://www.jstor.org/stable/41194032.
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., … Morey, R. D. (2017). Bayesian inference for psychology. Part II: Example applications with JASP. Psychonomic Bulletin & Review. doi:https://doi.org/10.3758/s13423-017-1323-7
Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., … Morey, R. D. (2017). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. Psychonomic Bulletin and Review, 1–23. http://doi.org/https://doi.org/10.3758/s13423-017-1343-3
Whalen, J., Gallistel, C. R., & Gelman, R. (1999). Nonverbal counting in humans: The psychophysics of number representation. *Psychological Science, 10*(2), 130–137. https://doi.org/10.1111/1467-9280.00120.
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.
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., … Dong, Q. (2012). Is order the defining feature of magnitude representation? An ERP study on learning numerical magnitude and spatial order of artificial symbols. PLoS One, 7(11). http://doi.org/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 nonnormal populations with unequal variances. *Canadian Journal of Experimental Psychology, 47*(3), 523–539. https://doi.org/10.1037/h0078850.
Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods, 12*(4), 399–413. https://doi.org/10.1037/1082-989X.12.4.399.