COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 48 (2024) e13409

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13409

# Modeling Magnitude Discrimination: Effects of Internal Precision and Attentional Weighting of Feature Dimensions

Emily M. Sanford,[a] Chad M. Topaz,[b] Justin Halberda[a]

[a]*Department of Psychological & Brain Sciences, Johns Hopkins University*
[b]*Department of Complex Systems, Williams College*

Received 25 January 2023; received in revised form 20 November 2023; accepted 17 January 2024

## Abstract

Given a rich environment, how do we decide on what information to use? A view of a single entity (e.g., a group of birds) affords many distinct interpretations, including their number, average size, and spatial extent. An enduring challenge for cognition, therefore, is to focus resources on the most relevant evidence for any particular decision. In the present study, subjects completed three tasks—number discrimination, surface area discrimination, and convex hull discrimination—with the same stimulus set, where these three features were orthogonalized. Therefore, only the relevant feature provided consistent evidence for decisions in each task. This allowed us to determine how well humans discriminate each feature dimension and what evidence they relied on to do so. We introduce a novel computational approach that fits both feature precision and feature use. We found that the most relevant feature for each decision is extracted and relied on, with minor contributions from competing features. These results suggest that multiple feature dimensions are separately represented for each attended ensemble of many items and that cognition is efficient at selecting the appropriate evidence for a decision.

*Keywords:* Approximate Number System; Magnitude discrimination; Psychophysics; Computational modeling; Generalized Magnitude System

When viewing a scene, multiple interpretations can often be supported by the same perceptual evidence. For instance, imagine that you are looking at a leafy green shrub on the edge of a forest path. The shrub is smaller than a nearby oak tree, but it has more leaves. So which plant should we say has “more?” Answering such a question requires a decision of what feature to focus on: physical size, or the number of leaves. How do we determine

Correspondence should be sent to Emily M. Sanford, Department of Psychology, University of California, Berkeley, 2121 Berkeley Way, Berkeley, CA 94720, USA. E-mail: esanford@berkeley.edu which interpretation is right in the moment, and how do we maintain that interpretation while suppressing conflicting ones?

Sometimes, perceptual evidence dramatically supports one interpretation over another, and low-level salience can lead to a predisposition toward one interpretation (Itti, Koch, & Niebur, 1998). Other times, our experience can spontaneously switch between multiple interpretations in the absence of differences in saliency, as is the case with bistable images (Kleinschmidt, Büchel, Zeki, & Frackowiak, 2002). Intentional selection can also guide which features to focus on, especially if the goal is known in advance (as in visual search tasks, e.g., "Decide as quickly as possible whether there is a green H among red Hs and green Xs"; Treisman, 1982), but also retrospectively (in memory search, e.g., "Was this item present in the previously-memorized list?"; Cavanagh, 1972). Perceptual evidence and goals interact in varied ways to determine what one focuses on.

Here, we use magnitude perception as a case-study to investigate how the mind navigates the process of selecting among multiple interpretations. The evolutionarily ancient ability to rapidly extract the number of objects from ensembles relies on the Approximate Number System (ANS; for a review, see Feigenson, Dehaene, & Spelke, 2004). Similarly, we have the ability to encode the convex hull that surrounds a cluster of items (Clayton & Gilmore, 2015; Clayton, Gilmore, & Inglis, 2015; Gilmore, Cragg, Hogan, & Inglis, 2016; Norris, Clayton, Gilmore, Inglis, & Castronovo, 2019) or the average or total area of items (Fuhs, McNeil, Kelley, O'Rear, & Villano, 2016; Gebuis & Reynvoet, 2011; Smets, Sasanguie, Szűcs, & Reynvoet, 2015; Szűcs & Myers, 2017; Szűcs, Nobes, Devine, Gabriel, & Gebuis, 2013). Our ability to discriminate ensembles based on these features is often studied using comparison tasks, where two groups of dots are displayed next to each other, intermixed, or in sequence, and the job of the participant is to respond which of the two groups contains more dots, a larger convex hull, or more total pixels (Anobile, Cicchini, Pomè, & Burr, 2017; Braham, Elliott, & Libertus, 2018; Dakin, Tibber, Greenwood, Kingdom, & Morgan, 2011; DeWind & Brannon, 2012; Feigenson et al., 2004; Franconeri, Bemis, & Alvarez, 2009; Fuhs et al., 2016; Gebuis, Cohen Kadosh, & Gevers, 2016; Halberda & Feigenson, 2008; Halberda, Mazzocco, & Feigenson, 2008; Libertus, Feigenson, & Halberda, 2013; Mazzocco, Feigenson, & Halberda, 2011; Odic, 2018; Odic, Hock, & Halberda, 2014; Odic, Libertus, Feigenson, & Halberda, 2013; Odic, Pietroski, Hunter, Lidz, & Halberda, 2013; Pica, Lemer, Izard, & Dehaene, 2004; Tomlinson, DeWind, & Brannon, 2020; Wang, Halberda, & Feigenson, 2017).

Number perception is a particularly interesting case for the question of how one selects among multiple interpretations of a stimulus because of the ongoing debate regarding the status of number representation. It has been argued that so-called "number responses" are actually responses to other ensemble features, such as surface area or convex hull (Clayton et al., 2015; Clayton & Gilmore, 2015; Dakin et al., 2011; Durgin, 1995, 2008; Gilmore et al., 2016; Morgan, Raphael, Tibber, & Dakin, 2014; Norris et al., 2019; Smets et al., 2015; Szűcs et al., 2013; Szűcs & Myers, 2017). Similarly, it has been claimed that number and surface area are perceived holistically as integral dimensions, and, therefore, cannot be separately represented (Aulet & Lourenco, 2021b). The primary evidence in support of shared underlying representations between number and non-numerical features comes in the form of congruency effects in discrimination tasks, where subjects are more accurate on trials where the larger group with respect to number is also the larger group with respect to other features, such as surface area (e.g., Hurewitz, Gelman, & Schnitzer, 2006). From an evolutionary perspective, it has been argued that shared representations between different ensemble features would be a useful strategy because features such as size and convex hull tend to be statistically correlated with number in the environment, thus making a concrete feature such as size a useful heuristic for number (Leibovich, Katzin, Harel, & Henik, 2017).

Others suggest that number is not only represented independently from other features, but in fact has a privileged representational status in the human mind (Anobile, Cicchini, & Burr, 2016; Clarke & Beck, 2021). Some evidence that number is represented independently of other ensemble features comes from cross-modal studies in which neonates match number across visual and auditory stimuli; these provide strong evidence that non-numerical visual features cannot be the only means through which number responses are generated (Izard, Sann, Spelke, & Streri, 2009). Also, approximate number performance predicts symbolic mathematical skills, providing evidence for shared *numerical* content between the representations underlying these different tasks (Halberda et al., 2008; Libertus, Feigenson, & Halberda, 2011; Wang et al., 2017). In response to congruency effects, it has been argued that, rather than reflecting shared underlying representations, such effects may instead occur due to Stroop-like response competition between independently represented dimensions (Clarke & Beck, 2021; Picon, Dramkin, & Odic, 2019).

With regard to the potential primacy of number, and consistent with it not being derived from representations of other ensemble features, in some studies, the congruency effect of number influencing judgments of area is stronger than the opposite effect (Tomlinson et al., 2020). Further, a tendency to spontaneously focus on numerosity over other ensemble magnitudes has been documented in both adults and young children, although there are individual differences in the degree of the effect (e.g., Cicchini, Anobile, & Burr, 2016; Hannula & Lehtinen, 2005). Compared to monkeys, humans are uniquely biased toward selecting number as their target dimension when categorizing groups of dots, and this bias is independent of mathematical education experience, again suggesting that number has a privileged status in the human mind (Ferrigno, Jara-Ettinger, Piantadosi, & Cantlon, 2017).

One way to partially reconcile these competing arguments is to posit shared computational resources underlying comparisons with different magnitudes, while maintaining that each magnitude is represented independently. For instance, the Shared Computations Account (SCA; Odic et al., 2013) claims that, while each magnitude dimension has its own dedicated representation—and thus its own internal precision—performance across dimensions is empowered by shared computational resources (e.g., shared ordinal, arithmetic, and logical computations; Odic et al., 2013). Such shared computation could result in correlated performance across dimensions despite the representations themselves having nonidentical precision. A nuance to the SCA account is that the required ordinal comparison machinery is typically assumed to be errorless in its execution and should, therefore, not lead to measurable individual differences (Halberda & Feigenson, 2008; Libertus et al., 2013; Odic et al., 2013; Piazza, Izard, Pinel, Le Bihan, & Dehaene, 2004, 2010). Thus, if we find correlations in performance across psychological dimensions, these may arise from shared executive function resources (e.g., staying on task), or error-prone ordinal comparisons (Brannon, Lutz,

& Cordes, 2006; Cordes & Brannon, 2008; Droit-Volet, Clement, & Fayol, 2008; Feigenson, 2007; Odic et al., 2013; Odic, 2018).

In this context, we sought to measure human flexibility at discriminating ensembles of dots based on three features: number, surface area, and convex hull. Although recent work indicates that people represent additive area rather than total surface area (Yousif & Keil, 2019; but see Park, 2022), the vast majority of previous number research has focused on total area (e.g., Clayton et al., 2015; DeWind, Adams, Platt, & Brannon, 2015; Leibovich et al., 2017; Norris et al., 2019; Tomlinson et al., 2020), so we decided to use this in our experiment; however, our experiment and model could easily be modified to accommodate for different an alternative area metric as well as for other features entirely. Participants were shown a stimulus set containing segregated groups of blue and yellow dots, and were instructed to respond, in separate blocks, which of the two groups was larger with respect to each feature. We chose total surface area and convex hull because they are some of the most-frequently investigated dimensions in the number literature (e.g., area: Brannon et al., 2006; DeWind, Bonner, & Brannon, 2020; Odic et al., 2013; Tomlinson et al., 2020; convex hull: Braham et al., 2018; Clayton et al., 2015; Clayton & Gilmore, 2015; Gilmore et al., 2016; Norris et al., 2019) and additionally, these dimensions can be varied separately from one another, which allowed us to quantify each of their individual contributions. Importantly, the stimuli were identical across the three tasks, which enables us to directly compare responses to a selfsame image. Subjects returned 3 months later to complete the experiment again, which allowed us to evaluate to what extent performance was stable over time.

Additionally, we introduce a computational model that empowers us to evaluate the extent to which subjects are relying on competing, nontarget features during a given task. A novel contribution of this model is that it allows us to take into account representational precision when quantifying the degree to which a given feature is being used. Although some theoretical work has been done to differentiate between these two constructs (e.g., Aulet & Lourenco, 2021a; Cicchini et al., 2016; Tomlinson et al., 2020), previous modeling efforts have typically conflated them (e.g., DeWind et al., 2015). An additional novel contribution of this work is that, with this modeling approach and our retesting procedure, we are able to investigate the stability not only of general performance on these magnitude comparisons, but also the internal factors (precision and feature reliance) underlying that performance.

We had four main behavioral predictions that we tested in this work. First, we expected to find differences in performance between tasks, particularly expecting surface area performance to be better than number performance (e.g., Odic et al., 2013). Importantly, with our model, we will be able to disentangle whether accuracy differences are due to differences in representational precision or differences in the relative weighting of target and nontarget features on each task. Second, we also expect to find congruency effects in all three tasks, where performance is worse on incongruent trials than on congruent trials. If congruency effects are indicative of shared representations underlying responses across different tasks, as has been claimed, then we would expect to find that the relative weightings of different features should be similar across tasks. Third, we expected to find stability in performance across time, particularly in number, which has been previously demonstrated and has been attributed to stability in representational precision (Clayton et al., 2015; DeWind &

Brannon, 2016; Elliott, Feigenson, Halberda, & Libertus, 2019; Libertus & Brannon, 2010; Price, Palmer, Battista, & Ansari, 2012; Purpura & Simms, 2018). This work will not only provide further evidence about the stability in magnitude comparison performance (including in other features, surface area and convex hull, that have not been as rigorously tested with respect to their stability), but will also provide a computational account for *why* performance is stable: is it due to stability in internal precision, feature weighting, or both? Finally, if number is privileged in the human mind relative to other ensemble features, we would expect to find that subjects rely more heavily on the number feature during the number task than they do on the other features during their respective tasks. Similarly, we would expect that number would be relatively highly weighted during the surface area and convex hull tasks (compared to the other nontarget feature).

## 1. Methods

### *1.1. Participants*

A total of 56 people participated in the study (self-reported gender: 43 females, 13 males). Participants were undergraduate students who participated for course credit. They ranged in age from 18 to 25 years old ($M = 19.75$, $SD = 1.39$ years). All participants had normal or corrected-to-normal vision and were not colorblind (determined by self-report; but note that stimuli differed in luminance and were spatially segregated, and thus colorblindedness was not a major concern).

### *1.2. Materials*

Stimuli were displayed on a Macintosh iMac computer monitor with a refresh rate of 60 Hz. The viewing distance was unconstrained but averaged approximately 57 cm and the display subtended $41.33 \times 26.02°$ of visual angle.

The stimuli consisted of 216 images depicting blue and yellow dots on a gray background (see Fig. 1). Blue dots were always presented on the left side of the image and yellow dots on the right. Images were generated using a custom algorithm in Python, where the number and total surface area (in pixels) of each dot set, as well as dot placement (which determined convex hull) could be specified, such that these features could be varied independently of one another. To vary the convex hull, dots were randomly placed within rectangular areas of varying sizes, and this allowed us to create smaller convex hulls by requiring the program to place all the dots within a smaller rectangular area. Dots were placed so that there were at least 10 pixels between neighboring dots and between each dot and the edge of the image. The exact value of each feature was then measured precisely after the stimuli were generated: the total surface area for an ensemble was quantified as the number of pixels of that color in the image, and the area of the convex hull of each set was precisely computed using the Convex Hull function in Python's SciPy package (Jones, Oliphant, & Peterson, 2001). Stimuli were iteratively generated using these two algorithms until the desired relationship between features across the entire stimulus set was reached.

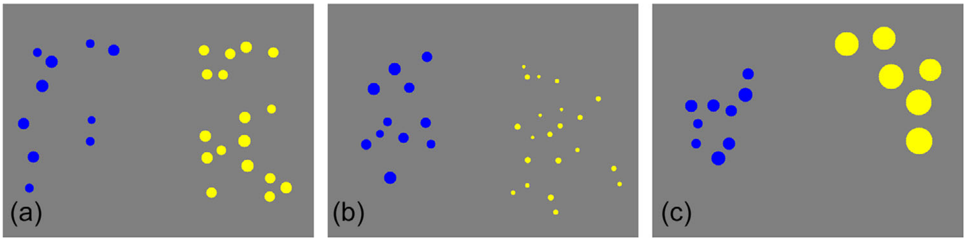

Fig. 1. Example stimuli, which varied in their congruency between number, surface area, and convex hull.
*Note*. (A) All three features are congruent (yellow is larger than blue for number, surface area, and convex hull). (B) Number and convex hull are congruent, but surface area is incongruent with them (yellow is larger than blue for number and convex hull, but blue is larger than yellow for surface area). (C) Both convex hull and surface area are incongruent with number (yellow is larger than blue for convex hull and surface area, but blue is larger than yellow for number).

Across the images, there was no difference in number, surface area, or convex hull values between the blue and yellow dot sets, all $t$s $< 1$, all $p$s $> .424$. The number of dots varied between 5 and 22 ($M = 12.91$, $SD = 5.30$). Total surface area ranged from 1784 to 19,188 pix ($M = 9702.95$, $SD = 4800.77$). Convex hull values ranged from 12,699 squared units (approximately contained within a box of size 89×164 pixels) to 104,174 squared units (approximately contained within a box of size 260×462 pixels; $M = 50{,}698.9$, $SD = 20{,}098.4$). These values were chosen based on the size of the stimulus images and pilot testing.

Intuitively, the difficulty of a trial is determined by the ratio of the relevant feature between the blue and yellow dot sets (larger value divided by smaller). In the number task, ratios varied from the hardest at 1.09 (e.g., 11 vs. 12) to the easiest at 2.86 (e.g., 7 vs. 20; $M = 1.87$, $SD = .49$). In the surface area task, ratios varied from the hardest at 1.09 (e.g., 4128 vs. 4485 pixels) to the easiest at 2.99 (e.g., 2801 vs. 8365 pixels; $M = 2.03$, $SD = .54$). In the convex hull task, ratios varied from the hardest at 1.01 (e.g., 57,981 vs. 58,540 units$^2$) to the easiest at 4.50 (e.g., 15,664 vs. 70,473.5 units$^2$; $M = 1.88$, $SD = .66$). In each task, yellow and blue were each the correct response to exactly half of the trials.

One of the main motivations for this research was to evaluate the extent to which a given feature influences performance even when that feature has *no predictive power* for the correct response. That is, are subjects *obligated* to use information from nontarget features, even when that information cannot improve their performance? In essence, we are interested in the "baseline" amount that nontarget features are used during a given task, which would occur when those features are completely not predictive of the correct response with the target feature. To test this, it was essential that our features be completely orthogonalized across the stimulus set and fully balanced for congruency. (Note, however, that the features being orthogonalized in the stimulus set certainly does not ensure that they are represented orthogonally in the mind, or that they are even being represented at all, as evidenced by the arguments surrounding additive vs. total area; see Yousif & Keil, 2019).

Therefore, we developed a novel method to orthogonalize competing features in the stimulus set. On a given trial, the "larger" group (yellow or blue) with respect to one feature, for example, convex hull, should have a 50% chance of also being the "larger" group for the other feature, number (i.e., there should be a 50% chance that number and convex hull are congruent), such that one could not "use" convex hull to respond correctly on the number task over the course of the experiment. To ensure this lack of predictive power, we generated our stimuli such that there were equal numbers of congruent and incongruent trials for each of the nontarget features: there were 108 stimuli where, for example, number and convex hull were congruent, and 108 stimuli where number and convex hull were incongruent. The same was true for number and surface area, as well as surface area and convex hull.

Additionally, we sought to make the difficulty of a given stimulus image with respect to one feature unrelated to the difficulty of that same stimulus image with respect to another feature. That is, there should be trials where the number comparison is easy and the convex hull comparison is hard, and vice versa, as well as trials where both are easy, and trials where both are hard; and these should all be equally distributed across congruent and incongruent trials. To evaluate whether this had been accomplished, we performed a correlation analysis to assess the relationships between feature ratios. We used transformed ratios, where we calculated each stimulus's ratio for each dimension by dividing the larger (yellow or blue) by the smaller for that dimension, then we subtracted one from each ratio (so that the point of equality between yellow and blue would be at a ratio of zero instead of one). If blue was larger than yellow, we then multiplied the ratio by $-1$, as this would allow us to account for both the *magnitude* of the ratio and the *direction* of correct response. With these transformed ratios, we were then able to examine the correlations between the different feature ratios. This analysis allowed us to conclude that the ratios of each feature were orthogonal with respect to the ratios of the other two features: Number and Surface Area ($R^2 = .001$), Number and Convex Hull ($R^2 = .020$), Surface Area and Convex Hull ($R^2 = .029$).

An additional benefit of our orthogonalization procedure is that it allows us to easily disambiguate the *influence* of each target feature. Since there is nearly no collinearity between, for example, the number and convex hull ratios, it is computationally possible to conclusively assess the effect of convex hull on number responding. This would not be possible if number and convex hull were at all correlated, as this would result in ambiguity about which feature was truly responsible for capturing the variance in responding. For example, this is the case in, for example, DeWind et al. (2015)'s stimulus space, where number and convex hull are partially correlated with one another.

### *1.3. Procedure*

Participants were tested individually in a quiet room by a trained experimenter. All participants saw the same set of 216 images three times, each time in a separate block, to compare the relative magnitudes of the blue and yellow dots with respect to each of the three features: number, surface area, and convex hull. The order in which the three tasks were presented was randomized across participants. Within a given task, all participants saw the images in the same order, and this order was created to ensure that one response ("Yellow" or "Blue") would not be the correct answer for more than three trials in a row. The group saw the images in a different, unique, order for each task.

Prior to each block, participants were shown an instruction screen that explained the task and gave examples to illustrate the relevant feature for that block. This was especially important prior to the convex hull task, where three example stimuli with white dashed lines drawn around the hull of the sets were shown to demonstrate the concept of convex hull. After the instruction screen, participants were given seven practice trials with feedback, and the experimenter walked through the first set of instructions and practice trials with the participants to ensure they understood the task.

Following the practice trials, the participant completed 216 experimental trials, with a break halfway through (after 108 trials). Each trial started with a white fixation cross presented at the center of the screen for 280 ms. The stimulus was then presented for 250 ms, followed by a yellow and blue pixelated mask. The mask remained on the screen until a response was recorded. After the participants responded, they saw a blank screen for 800 ms before the next trial started. Participants were not given feedback on their performance during the experimental trials, and their response, as well as response time, were recorded.

### *1.3.1. Retesting*

Of our 56 participants, 40 returned to participate in the study a second time, approximately 3 months later. The experimental setup and design were identical between the two instances, and the order in which they completed the three tasks was randomized independently each time.

## 2. Computational model

Previous ANS research has extensively modeled the relationship between performance and internal precision (e.g., Halberda & Feigenson, 2008; Libertus et al., 2013; Odic et al., 2013; Piazza et al., 2004, 2010). In particular, a widely used function assumes that a subject's Weber Fraction, $w$, can be estimated based on the following function relating performance ($p$) and the ratio of the stimuli being compared ($r$), as well as a lapse parameter ($g$):

$$f_1\left(w, g, r\right) = p\left(correct\right) = \left(1 - g\right) * \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{r-1}{w\sqrt{2}\sqrt{1+r^2}}\right)\right] + \frac{g}{2}$$

Notice that this model assumes that the only information the subject retrieves from the stimulus is the ratio of the compared groups on the dimension on which they are being compared. We were interested in quantifying the extent to which nontarget feature information could influence performance, which was not previously accounted for in typical models (e.g., Halberda et al., 2008). An important element that we wanted to include in our model was that, if a nontarget feature such as convex hull were to influence responses on the number task, the amount of influence it exerted would be dependent on *how precise* that subject's representations of both the target and nontarget features are.

An example of the implications of this distinction is as follows. High reliance on a *highly precise* nontarget feature representation would result in *consistently incorrect* responses on trials where the correct response for the target and nontarget feature are incongruent, and consistently correct responses on trials where they are congruent. In contrast, high reliance on a highly *imprecise* nontarget feature representation would lead to closer to chance responding across all trials, with much less of a difference in performance based on congruency. Without separating precision and reliance, the latter situation would instead be attributed to *less reliance* on the nontarget feature than the first case, even if the representations were being relied upon equally in the two cases.

This distinction contrasts with previous models attempting to quantify the contribution of non-numerical features during a number task, such as DeWind et al. (2015). In DeWind et al.'s model, which regresses choice responses over orthogonal dimensions *Number*, *Size* (related to area), and *Spacing* (related to convex hull), the amount that a subject relied on the number dimension during a number task is quantified by the value of the regressor $\beta_N$. Notably, however, precision was derived from the same regressor ($w_N = \frac{1}{\sqrt{2}\beta_N}$). Therefore, this model is unable to *separately* quantify precision and amount of use (i.e., it cannot distinguish between a highly precise representation that is given a low weight, versus an imprecise representation that is given a high weight). We propose a new modeling approach to disentangle these two constructs and evaluate their separability.

In our model, the probability that a subject responds correctly on a given trial is a function of the precision of their representation for each dimension, as well as the amount that they rely on that particular dimension. For simplicity, we assume that the subject only represents and is influenced by three features of the stimuli (number, convex hull, and surface area), in particular the ratios between the two ensembles depicted on a given trial ($r_N$, $r_{CH}$, and $r_{SA}$). We also assume they have a separate value of internal precision for each feature ($w_N$, $w_{CH}$, and $w_{SA}$, respectively). Note, however, that this model can be expanded to include *any* feature of interest from the stimulus; these three were chosen because they can be independently orthogonalized, and because of their prevalence in the number perception literature.

During a trial of the number task, we assume that the subject intends to respond solely on the basis of number, yet they cannot help but extract and represent the ratio of convex hull and surface area between the two groups as well, subject to their precision for those features. Therefore, when they make their response, their probability of responding correctly is not only a function of $w_N$ and $r_N$, but also $w_{CH}$ and $r_{CH}$, as well as $w_{SA}$ and $r_{SA}$. We assume that one's representational precision for a given feature (e.g., $w_N$) is constant no matter what task is being performed. Finally, we assume that the amount that they are influenced by each feature (denoted here by $b$) varies, with the constraint that $b_N + b_{CH} + b_{SA} = 1$. This parameter indicates the relative weight given to that feature during the task.

Finally, to account for the fact that on congruent trials, information from a nontarget feature should *increase* the likelihood of responding correctly, whereas on incongruent trials, it should *decrease* the likelihood of responding correctly, we include a binary variable, $a$, for each feature ($a_N$, $a_{CH}$, $a_{SA}$), which indicates whether that feature is congruent ($a = 1$) or not ($a = 0$) with the target feature. On the number task, $a_N$ always equals 1, while $a_{CH}$ and $a_{SA}$ equal 1 and 0 on exactly half of the trials each in our stimuli. As a result, when the nontarget feature is incongruent with the target feature, the probability of responding correctly derived from the nontarget feature is *subtracted* from one, decreasing the overall likelihood of a correct response.

Then, the probability that the subject would correctly assess which group is larger on that trial is calculated according to the following functions, where *b*s, *w*s, and *g* (bolded) are fitted parameters:

$$f_N = a_N * [\boldsymbol{b_N} * f_1(\boldsymbol{w_N}, \boldsymbol{g}, r_N)] + (1 - a_N) * [1 - (\boldsymbol{b_N} * f_1(\boldsymbol{w_N}, \boldsymbol{g}, r_N)\Big]$$

$$f_{SA} = a_{SA} * [\boldsymbol{b_{SA}} * f_1(\boldsymbol{w_{SA}}, \boldsymbol{g}, r_{SA})] + (1 - a_{SA}) * [1 - (\boldsymbol{b_{SA}} * f_1(\boldsymbol{w_{SA}}, \boldsymbol{g}, r_{SA})\Big]$$

$$f_{CH} = a_{CH} * [\boldsymbol{b_{CH}} * f_1(\boldsymbol{w_{CH}}, \boldsymbol{g}, r_{CH})] + (1 - a_{CH}) * [1 - \boldsymbol{b_{CH}} * f_1(\boldsymbol{w_{CH}}, \boldsymbol{g}, r_{CH})]$$

The resulting values are then added together to get a final sum predicting percent correct on that trial:

$$p_{total} = f_N + f_{SA} + f_{CH}$$

Note that, because we hypothesized that subjects use different features to different degrees during the different tasks, we create *three* separate parameters for each of $b_N$, $b_{CH}$, and $b_{SA}$ (i.e., number task $b_N = / =$ surface area $b_N$). If the weight with which subjects use each feature does not vary between tasks, our model would assign a similar value for each *b* for each task. As a result, this model fits 13 parameters for each subject.

A perfect observer would only use number information during the number task, convex hull information during the convex hull task, and area information during the area task—completely ignoring the nontarget features (e.g., in the number task, $b_N = 1$, $b_{CH} = 0$, and $b_{SA} = 0$). On the other hand, an observer who uses a feature-neutral or generalized magnitude strategy might equally rely on all three features (i.e., $b_N = 0.33$, $b_{CH} = 0.33$, and $b_{SA} = 0.33$).

Now that we have our model delineated, we can use it to concretely illustrate the problem that is created when precision and feature weighting are conflated. Consider Fig. 2, which illustrates the relationship between weight, precision, and percent correct (denoted by color) for a single trial of the number task. On this toy trial, the number ratio is 1.57 and the convex hull ratio is 1.15, and these two features are incongruent ($a = 0$); for this trial, it would be harder to compare convex hull than number if the representational precision of the two features were equal.

The important point about this plot is that similar probabilities of responding correctly (e.g., positions A and B) emerge across large swaths of the graph, despite stark differences in the underlying parameters. Consequentially, a model that infers only *precision* from response data (e.g., DeWind et al., 2015) would assign the same precision to individuals that, in actuality, *greatly* differed in their internal precision (due to their underlying differences in weight, *b*). When fitting our model, we gain *independent purchase* on precision and weight by fitting these independent parameters across a range of stimulus values, across three separate tasks.

Given that internal precision has been found to relate to higher-level mathematical abilities in models that did not account for feature reliance (e.g., Halberda et al., 2008; Starr, Libertus, & Brannon, 2013; Wang et al., 2017), it is important to investigate which component is

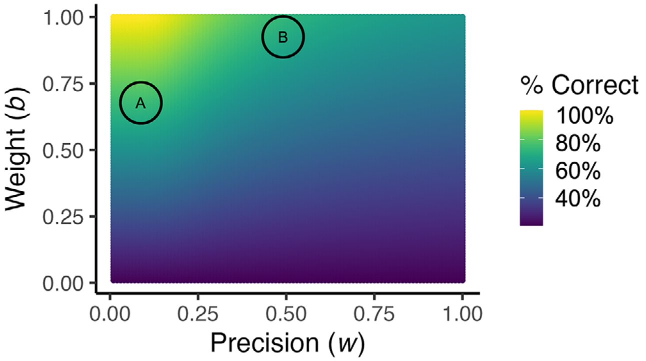

Fig. 2. Effect of target feature precision and reliance on expected percent correct for one particular trial of the number task.

*Note*. Both parameter combinations A and B yield a 69.1% chance of responding correctly on this number trial, despite the underlying parameters differing greatly (A: $w_N = .09$, $b_N = .69$; B: $w_N = .50$, $b_N = .95$).

actually responsible for the relationship. Differentiating between these two constructs will allow us to better understand the cognitive consequences of internal precision.

### 2.1. *Model validation*

To empirically ensure that the model can accurately recover parameters in data, we ran a series of simulations (see Supplementary Materials A for details).

## 3. Results

### 3.1. *Data cleaning*

For each task for each participant, we removed Response Time (RT) outliers that were at least 3 standard deviations away from their mean RT for that task (1.7% of trials; for an analysis on the effect of nontask features on RT, see Supplementary Materials B). Then, we calculated each subject's accuracy on each task at each time point, and excluded participants whose accuracy was greater than 3 standard deviations below the mean accuracy across all subjects for that task. This resulted in the removal of four subjects (one from the convex hull task at T1, one from the convex hull task at T2, and two from the surface area task at T2). After exclusions, there were 52 subjects with data at T1, and 40 of those subjects also had data at T2. Therefore, all results based on comparisons between time points only include those 40 subjects.

### 3.2. *Overall performance: Best on convex hull and varying congruency effects*

In terms of overall accuracy, subjects performed best on the convex hull task ($M = 90.7\%$, $SD = .05\%$), next best on the surface area task ($M = 88.3\%$, $SD = .04\%$), and worst on the number task ($M = 84.2\%$, $SD = .05\%$). A repeated measures one-way ANOVA confirmed that accuracy differed by task, $F(1.69,66) = 28.61, p < .001$, and follow-up $t$-tests with Bonferroni corrections confirmed that all three tasks differed from one another ($p$s $< .001$). Superior performance on surface area discrimination compared to number discrimination is consistent with prior work (Odic et al., 2013).

Next, we investigated whether performance on one task predicted performance on the other tasks. Correlation tests using Bonferroni corrections for multiple comparisons revealed that subjects who performed more accurately on the number task also performed more accurately on the surface area task, $r(90) = .333$, $p = .001$. A similar relationship was found between surface area and convex hull performance, $r(90) = .370$, $p < .001$. However, there was no relationship in accuracy between the number and convex hull tasks, $r(90) = -.028, p = .791$. Thus, we found weak evidence for shared ability between the different tasks.

As expected, subjects tended to perform better on trials where nontarget features were congruent with the target feature. This was confirmed with repeated measures $t$-tests with Bonferroni corrections. On the number task, average performance was higher on trials where surface area was congruent with number ($M = 85.1\%$, $SD = .05\%$) than on incongruent trials ($M = 82.5\%$, $SD = .06\%$), $t(51) = 3.62$, $p = .004$, and this was even more dramatically evident for convex hull congruency (congruent: $M = 95.4\%$, $SD = .03\%$; incongruent: $M = 71.6\%$, $SD = .10$), $t(51) = 15.89$, $p < .001$. On the surface area task, again, convex hull congruency significantly impacted performance (congruent: $M = 92.8\%$, $SD = .05\%$; incongruent: $M = 82.3\%$, $SD = .07\%$), $t(51) = 9.66$, $p < .001$. Interestingly, on the surface area task, subjects were slightly better on trials where number was *incongruent* ($M = 88.9\%$, $SD = .08\%$) than congruent ($M = 86.5\%$, $SD = .05\%$), although this difference was not significant, $t(51) = 1.89, p = .39$. On the convex hull task, subjects performed better on trials where number was congruent ($M = 93.7\%$, $SD = .04\%$) than incongruent ($M = 85.6\%$, $SD = .07\%$), $t(51) = 8.88$, $p < .001$, and the same was true for surface area (congruent: $M = 91.8\%$, $SD = .04\%$; incongruent: $M = 87.5\%$, $SD = .06\%$), $t(51) = 5.77, p < .001$.

In some cases, these congruency effects are extremely strong (e.g., a 23.8% difference in number performance between convex hull congruent and incongruent trials). Nonetheless, subjects remained above chance even on incongruent trials, indicating that they were still more likely to give the target response than a nontarget response. This means that subjects provided different responses to the very same image when asked to compare on the basis of different (conflicting) features. We confirmed this with a series of logistic regressions predicting responses on incongruent trials of one task from responses to the same image during the alternate task. We found significant *negative* relationships in all three task combinations: when predicting number responses from surface area responses ($B = -.88$, $p < .001$, odds ratio = 12.7%); when predicting number responses from convex hull responses ($B = -2.06$, $p < .001$, odds ratio = 41.4%), and when predicting surface area responses from convex hull responses ($B = -.53, p < .001$, odds ratio = 58.9%). Subjects were more likely to respond *differently* on two tasks if the features were incongruent with one another. Although they performed worse overall on incongruent trials than on congruent trials, subjects nonetheless seem to be primarily responding on the basis of the target feature, even when it is incongruent with the other features.

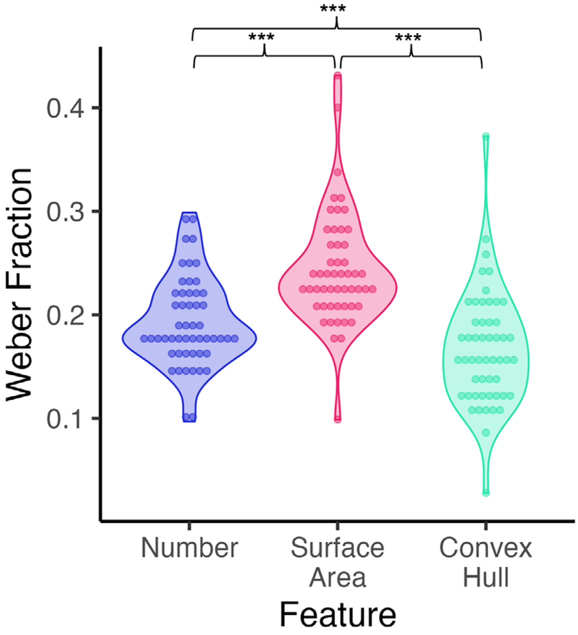

Fig. 3. Modeled Weber Fractions for each feature.
*Note*. Lower Weber Fractions correspond to better acuity.

### 3.3. *Modeling results*

For our first series of analyses, we fit our model using Maximum Likelihood Estimation (MLE) to each subject's total dataset (i.e., collapsed across time if the subject had data from both T1 and T2).

#### 3.3.1. *Number is more precise than surface area, less than convex hull*

First, we asked whether acuity ($w$) was similar for the three features. We found significant differences in modeled precision across the three tasks using a one-way ANOVA, $F(2,152) = 31.92$, $p < .001$ (see Fig. 3). With follow-up $t$-tests using Bonferroni corrections, we found that subjects were significantly more precise on the convex hull task, as indicated by lower values of $w_{CH}$ ($M = .165$, $SD = .056$), than $w_N$ ($M = .194$, $SD = .042$), $p = .017$, which was in turn significantly more precise than $w_{SA}$ ($M = .244$, $SD = .054$), $p < .001$. This result was consistent with convex hull being the most accurate of the three tasks. Interestingly, although subjects were on average more *accurate* on the surface area task than the number task, we found that their modeled *precision* patterned in the opposite direction.

Next, we asked whether precision patterned together across the three tasks. That is, does the precision of a person's numerical representations predict the precision of their surface area or convex hull representations? Correlation tests with Bonferroni corrections for multiple comparisons revealed that while $w_N$ correlated with $w_{SA}$, $r(50) = .425$, $p = .002$, and $w_{CH}$ patterned with $w_{SA}$, $r(50) = .405$, $p = .003$, there was no relationship between $w_{CH}$ and $w_N$,

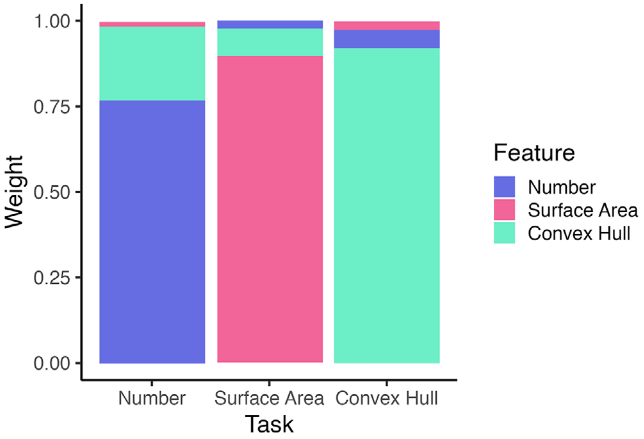

Fig. 4. Average weights assigned to each feature during each task.

$r(50) = .098$, $p = .488$. This mixed result provides inconclusive evidence with respect to a potential linkage in precision across different magnitudes. Given that we have no a priori reason to expect that number and convex hull representations should be less closely linked than other combinations of our features, future work could investigate whether this result is consistent in other samples and, if so, why.

### *3.3.2. Subjects rely on different features for different tasks*

We then turned to investigate the weight ($b$) distributions across each task. First, we asked whether subjects used the same combination of weights regardless of task—that is, was $b_N$ during the number task similar to $b_N$ during the surface area task? A two-way ANOVA predicting weight from feature and task revealed that subjects used very different distributions of weights during each task. Most importantly for our question, there was a strong interaction between feature and task, $F(4,458) = 2368.96$, $p < .001$ (see Fig. 4).

This interaction effect indicates that, for example, the amount that subjects used their number representation depended on whether they were performing the number task or not. Subjects in fact weighted the target feature most strongly for each of the three tasks (number task $b_N$: $M = .773$, $SD = .127$; surface area task $b_{SA}$: $M = .897$, $SD = .100$; convex hull task $b_{CH}$: $M = .920$; $SD = .061$). On both the number and surface area tasks, the second most highly weighted feature was convex hull (number task $b_{CH}$: $M = .213$, $SD = .129$; surface area task $b_{CH}$: $M = .080$, $SD = .087$). On the convex hull task, number was the second most highly weighted feature (convex hull task $b_N$: $M = .056$, $SD = .057$). Interestingly, this means that surface area was the *least* weighted nontarget feature on both the number and the convex hull tasks (number task $b_{SA}$: $M = .014$, $SD = .024$; convex hull task $b_{SA}$: $M = .024$, $SD = .040$), indicating that surface area was not strongly influencing responses during the other tasks. One possible reason for this is that our subjects may be actually representing additive area rather than total surface area, such that our model fails to capture weight placed on the similar, but nonidentical, area metric (Yousif & Keil, 2019). Notably, the weight on number during the number task was significantly lower than either the weight for surface area during the surface area task or the weight for convex hull during the convex hull task, $p$s $< .001$, while the latter two did not differ from one another, $p = .16$. This indicates that subjects had a relatively harder time *focusing* on number during the number task than on the other two features during their respective tasks. This result is inconsistent with the claim that number has a privileged position as a represented dimension in the human mind.

Next, we asked whether the ability to focus on the target feature was a domain-general ability. That is, are subjects who are the best at focusing on number (high $b_N$) during the number task also the best at focusing on convex hull (high $b_{CH}$) during the convex hull task, perhaps because they have better general inhibitory control than other participants? In fact, we found the opposite: a correlation between number-task $b_N$ and convex hull-task $b_{CH}$ yielded a *negative* relationship, $r(50) = -.354$, $p = .010$. There was a positive relationship between number-task $b_N$ and surface area-task $b_{SA}$, $r(50) = .370, p = .007$, and no relationship between surface area-task $b_{SA}$ and convex hull-task $b_{CH}$, $r(50) = .153$, $p = .280$. These inconclusive results indicate, at a minimum, that the ability to home in on the target feature is *not* best thought of as a domain-general inhibitory ability, and may instead be consistent with prior work suggesting the existence of *number-specific* inhibitory control (e.g., Piazza, De Feo, Panzeri, & Dehaene, 2018; Wilkey & Price, 2019).

A possible explanation for this last pattern of results would be that subjects have a more rigid weight distribution across the three tasks. For example, it is possible that a subject who relies highly on convex hull during the convex hull task (high $b_{CH}$) would have a harder time *suppressing* convex hull during the number task (again, high $b_{CH}$), and as a result, would have a correspondingly *lower* weight on number during the number task (low $b_N$). Therefore, we evaluated whether there was a relationship between target feature weight, and weight for that feature on the other tasks (i.e., number-task $b_{CH}$ vs. convex hull-task $b_{CH}$). This hypothesis was somewhat substantiated by the positive relationship between convex hull-task $b_{CH}$ and number-task $b_{CH}$, $r(50) = .403$, $p = .003$, which indicates that subjects who used convex hull relatively more during the convex hull task also tended to use it relatively more during the number task. Likewise, we found that there was a positive relationship between number-task $b_N$ and convex hull-task $b_N$, $r(50) = .348$, $p = .012$, indicating that subjects who used number more during the number task also used number more during the convex hull task. In contrast, there was no relationship between convex hull-task $b_{CH}$ and surface area-task $b_{CH}$, $r(50) = .086$, $p = .542$, nor between number-task $b_N$ and surface area-task $b_N$, $r(50) = .135$, $p = .341$. Similarly, there was no relationship between the surface area-task $b_{SA}$ and either the number-task $b_{SA}$, $r(50) = .025$, $p = .863$, or convex hull-task $b_{SA}$, $r(50) = -.032$, $p = .820$.

This provides some support for the idea that subjects who use a feature more during its target task also use it relatively more—or have a harder time suppressing it—when that feature is *not* the target. However, this only appears to be the case for features that are playing a relatively larger role in their estimates, as this pattern was not seen with surface area, which was the least used nontarget feature.

### *3.3.3. Subjects do not rely the most on the most precise representations*

As previously stated, the main motivation for this research was to tease apart the constructs of precision and reliance. If they are actually related with one another, we would see a negative correlation between weight and reliance (where subjects with better precision are using the feature relatively more). This could even be conceived as an optimal way to combine information with the highest fidelity: if subjects are sensitive to their own precision, they should more highly weight their more precise representations and rely less upon their less precise representations, since they can be less sure that those contain accurate information about the world. To the extent that feature weights are tuned to longer-term priors (e.g., "When I estimate number in the world, I find it helpful to weight convex hull somewhat highly"), then we should observe some of this up-weighting across tasks.

Some preliminary evidence against this conclusion can be gleaned from the previous analyses. For instance, on average, subjects have more precise number than surface area representations ($w_N < w_{TSA}$), yet are on average weighting surface area on the surface area task relatively *more* than they are weighting number on the number task (number-task $b_N <$ surface area-task $b_{TSA}$). This is not optimal, since they should be using number *more* relative to surface area if their number representations are more precise. Given our orthogonalization of feature dimensions, relying on surface area in the number task cannot aid accurate decisions.

Additionally, we directly investigated whether subjects with a more precise representation (lower $w$) were using that feature more within that task (higher $b$) relative to other subjects (i.e., $w$ and $b$ should be negatively correlated across subjects). We found that this was not the case for any feature (number: $r(50) = .227$, $p = .106$, $\mathrm{BF}_{10} = 1.022$; surface area: $r(50) = -.078$, $p = .584$, $\mathrm{BF}_{10} = .358$; convex hull: $r(50) = .025$, $p = .862$, $\mathrm{BF}_{10} = .317$). We found no evidence that subjects who had more precise representations weighted those representations more highly on the target task relative to other subjects. In fact, two of the trends were in the opposite direction, where subjects with *worse* precision were using that feature slightly *more*. Resultingly, feature precision and reliance appear to be separable constructs that do not pattern together during magnitude comparison tasks.

## *3.4. Comparing results across time*

Using the data from the 40 subjects who returned to complete the study again after 3 months, we were able to investigate whether performance in each task is stable across time, and to compare within- and across-task stability.

First, we investigated whether accuracy (% correct) on a given task at T1 predicted accuracy on the *same task* at T2, consistent with prior research. We found that it did, robustly, for all three tasks (Number: $r(38) = .690, p < .001$; Surface Area: $r(38) = .649, p < .001$; Convex Hull: $r(38) = .535, p < .001$).

Split by time, we found less robust evidence of related performance *across* tasks. Comparing all three tasks within each time point, we found that there were significant positive relationships only between accuracy on the convex hull and surface area tasks at T1, $r(38) = .580, p < .001$, and between the number and surface area tasks at T2, $r(38) = .476, p = .002$.

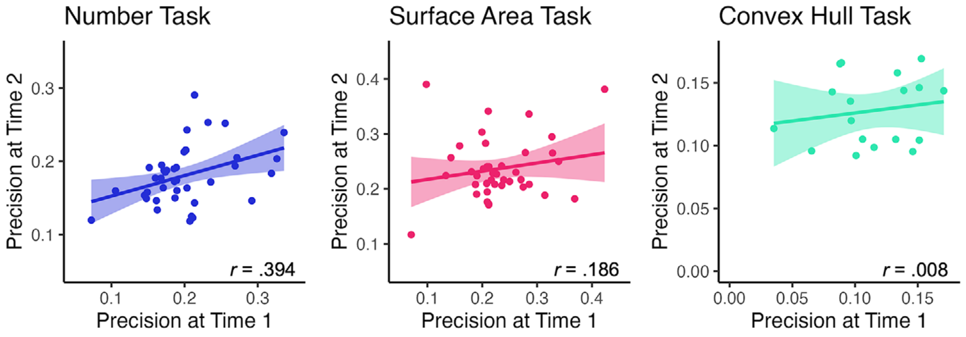

Fig. 5. Stability in precision over time.

No other comparisons were significant, *ps* > .163. Overall, accuracy does not appear to be strongly related across the three tasks.

### *3.5. Modeling results across time*

We once again fit our model to each subject's data, this time separately for T1 and T2.

#### *3.5.1. T1 and T2 yielded very similar modeling results*

First, we asked whether results from the time-collapsed models (e.g., differences in precision on the three features) were driven differentially by performance at either time point. A two-way (feature × time) repeated measures ANOVA revealed that there was again a significant difference in precision between the three features, $F(2{,}233) = 48.94$, $p < .001$. Once again, $w_{CH}$ ($M = .149$, $SD = .051$) was more precise than $w_N$ ($M = .191$, $SD = .048$), $p < .001$, which was in turn more precise than $w_{SA}$ ($M = .234$, $SD = .062$), $p < .001$. There was no effect of time on precision, $F(1{,}233) = .018$, $p = .894$. There was also a marginal interaction between task and time, $F(2{,}233) = 1.64$, $p = .196$. Overall, performance was similar on a given task between the two time points.

#### *3.5.2. Number precision is stable over time*

Previous research has attributed temporal stability in number performance to stability in one's representational precision (Clayton et al., 2015; DeWind & Brannon, 2016; Elliott et al., 2019; Libertus & Brannon, 2010; Price et al., 2012; Purpura & Simms, 2018). However, these previous models have not included the feature weighting component, which is another possible locus of stability in performance. Therefore, we investigated whether acuity ($w$) at T1 predicted acuity at T2 for any of our three features. Consistent with previous research, we found a significant relationship between number precision at T1 and T2, $r(38) = .394$, $p = .012$. In contrast, neither surface area nor convex hull showed significant stability in precision (Surface Area: $r(38) = .186$, $p = .252$; Convex Hull: $r(38) = .008$, $p = .959$; see Fig. 5).

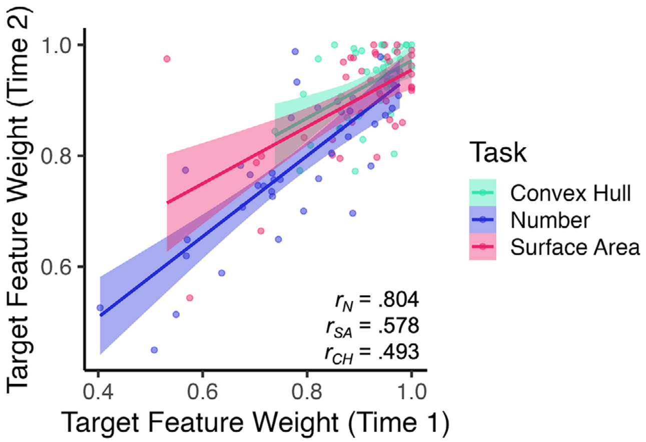

Fig. 6. Stability in weights over time.

### 3.5.3. *Feature weighting is highly stable over time*

The result that neither surface area nor convex hull precision were stable across time was surprising considering that we found a strong relationship between *accuracy* on each task between T1 and T2. Therefore, we turned to investigate whether feature *weightings* were stable across time for each task (e.g., does surface area-task $b_{SA}$ at T1 predict surface area-task $b_{SA}$ at T2?). Interestingly, we found strong evidence of target feature weight stability for all three tasks (Number: $r(38) = .804$, $p < .001$; Surface Area: $r(38) = .578$, $p < .001$; Convex Hull: $r(38) = .493$, $p = .001$; see Fig. 6). This indicates that the ability to correctly focus on for example, number during the number task, while distinct from the ability to focus on convex hull during the convex hull task, is stable across time. This appears to be an additional source of stability in feature-specific magnitude comparison performance (and perhaps even the main source of stability for the surface area and convex hull tasks).

However, it is worth noting again that in our model validation (see Supplementary Materials A, *Stability in w, b*), when both precision and weight were correlated between T1 and T2, the simulation recovered a high correlation between weight at T1 and T2, but a much lower correlation between precision at T1 and T2. This may somewhat explain the lack of correlations we found in surface area and convex hull precision. Our results strongly suggest that feature reliance is a stable individual difference, while the results on the stability of internal precision are less clear. Further research is necessary before one can draw a firm conclusion about the stability of these different components of performance.

## 4. Discussion

In this study, we asked to what extent people combine the same information in different amounts depending on the task at hand. When investigating feature reliance as defined in our model, we found that the ability to focus on one feature does not predict the ability to focus on another feature. This indicates that the ability to focus on a particular dimension is not best conceived of as a domain general “inhibitory control” ability. In fact, our data trended in the opposite direction, such that high reliance on convex hull during the convex hull task predicted higher reliance on convex hull during the number task as well. These results are more consistent with a proposed number-specific inhibitory control ability (e.g., Piazza et al., 2018; Wilkey & Price, 2019). Our results are also consistent with prior studies that found convex hull mattered more than total area during number comparison (e.g., Clayton & Gilmore, 2015; Gilmore et al., 2016).

We also investigated the relationship between internal precision and reliance on target and nontarget features, as they had been previously conflated in approximate number models (e.g., DeWind et al., 2015). We found that these two parameters were uncorrelated in all three tasks, which indicates that a person’s internal precision and their ability to focus on a particular feature (and suppress other features) should be conceived of as separate contributors to overall magnitude comparison performance. Particularly because these dimensions are separable, future investigations of the influence of continuous features on number performance should take into account the representational precision of the other features, which will give a more accurate estimate of how much that feature is actually influencing number responses.

Interestingly, our results conflict with past research that has suggested that number has a privileged role in the human mind (e.g., Cicchini et al., 2016; Ferrigno et al., 2017; Tomlinson et al., 2020). In our results, number was not consistently the most-used nontarget feature on other comparisons, and it was the least strongly relied-upon feature on its own task (compared to the weight on surface area during the surface area task, or the weight on convex hull during the convex hull task). Instead, consistent with recent work indicating that number may not have a privileged role once feature perceptibility is taken into account (Aulet & Lourenco, 2023), these results suggest that number is not more easily focused on when other features conflict with it.

One of the novel contributions of this work is that we investigated the *source* of stability in ANS performance (e.g., Clayton et al., 2015; DeWind & Brannon, 2016; Elliott et al., 2019; Price et al., 2012; Purpura & Simms, 2018). Here, we have taken the previously stable $w$ measure and essentially partitioned it into two separate parameters. We found a significant relationship between $T_1$ and $T_2$ for $w_N$, consistent with previous research. In contrast, there was no evidence of stability in $w_{SA}$ or $w_{CH}$ between $T_1$ and $T_2$. We additionally found robust evidence of stability in *feature weighting* between $T_1$ and $T_2$, for every task. With the caveat that our model might be biased toward this conclusion (i.e., weight is stable and precision is not; see Supplementary Materials), it appears to be the case that, at least, the ability to focus on a target feature for a given task is stable over time, even while that ability is *not stable within a subject across tasks*.

Further research is necessary to more conclusively determine which of these components are stable across time, and why the ability to focus on one feature does not predict the ability to focus on another. Another question of interest is how domain-general inhibitory control abilities relate to the feature weightings we found in our current study. Although our three target-feature weights did not correlate with one another, there remains a possibility that some (or all) of these abilities relate at least partially with other measures of inhibitory control (such as number-specific inhibitory control; Piazza et al., 2018; Wilkey & Price, 2019). Finally, another question is how these separable components of magnitude comparison develop. The ANS improves across development, peaking in performance around age 30 (Halberda, Ly, Wilmer, Naiman, & Germine, 2012). Do developmental improvements in internal precision and target feature weighting equally contribute to this general improvement in performance? What are the trajectories of development for each of these components of performance? Some previous work suggests that the development of the ability to focus on number could be the key reason that ANS performance improves with age (e.g., Piazza et al., 2018); our modeling framework would allow for an explicit test of this hypothesis.

Parsing the previously singular parameter $w$ into two separate, independent parameters opens important questions for the field of numerical cognition. Particularly, it will be important to investigate the relationship between ANS performance and formal mathematics, where internal precision has been shown to predict mathematics performance (Halberda et al., 2008; Mazzocco et al., 2011). One possibility is that feature weighting (i.e., the ability to focus on number during the number task) may play a part in the relationship found in previous research.

Another valuable direction for future inquiry is to evaluate the developmental trajectories of precision and reliance. It has previously been shown that ANS precision improves between birth and age 30 (Halberda et al., 2012), but it is unknown to what extent that maturation is due to changes in *internal precision* versus changes in *the ability to focus on the correct feature*. Similarly, it may be valuable to quantify which abilities are more strongly affected in populations with ANS-specific deficits, such as dyscalculia and Williams' Syndrome (Butterworth, 2010; Castaldi, Turi, Gassama, Piazza, & Eger, 2020; Cheng et al., 2020; Dowker & Kaufmann, 2009; Libertus, Feigenson, Halberda, & Landau, 2014; O'Hearn & Luna, 2009; Piazza et al., 2010; Wilson, Revkin, Cohen, Cohen, & Dehaene, 2006), as it could help inform the best approaches for future numerical interventions.

Finally, although we only looked at three features here (number, surface area, and convex hull), the same modeling approach could be used for any combination of features, such as density, perimeter, or average area. Indeed, it could be applied to *any* magnitude comparison experiment with multiple competing dimensions. Using this delineated approach will allow us to better understand to what extent these other features influence perception and decision-making.

The picture of magnitude perception and judgment that emerges from this work is that each magnitude has its own precision in perception—there is not just one magnitude system; that each magnitude has its own weighting parameter; and these parameters are intelligently modulated for each task, and are stable across time in individuals.

## Acknowledgments

Data collection and analysis were supported by an NSF GRFP, DGE1746891, awarded to E.M.S. and a McDonnell Foundation Scholar Award awarded to J.H. We have no conflict of interest to disclose.

## Open Research Badges

This article has earned Open Data badge. Data is available at https://osf.io/xwqrz/.

## References

Anobile, G., Cicchini, G. M., & Burr, D. C. (2016). Number as a primary perceptual attribute: A review. *Perception*, *45*(1–2), 5–31. https://doi.org/10.1177/0301006615602599

Anobile, G., Cicchini, G. M., Pomè, A., & Burr, D. C. (2017). Connecting visual objects reduces perceived numerosity and density for sparse but not dense patterns. *Journal of Numerical Cognition*, *3*(2), 133–146. https://doi.org/10.5964/jnc.v3i2.38

Aulet, L. S., & Lourenco, S. F. (2021a). The relative salience of numerical and non-numerical dimensions shifts over development: A re-analysis of Tomlinson, DeWind, and Brannon (2020). *Cognition*, *210*(January), 104160. https://doi.org/10.1016/j.cognition.2021.104610

Aulet, L. S., & Lourenco, S. F. (2021b). Numerosity and cumulative surface area are perceived holistically as integral dimensions. *Journal of Experimental Psychology: General*, *150*(1), 145–156. https://doi.org/10.1037/xge0000874

Aulet, L. S., & Lourenco, S. F. (2023). No intrinsic number bias: Evaluating the role of perceptual discriminability in magnitude categorization. *Developmental Science*, *26*(2), 1–17. https://doi.org/10.1111/desc.13305

Braham, E. J., Elliott, L., & Libertus, M. E. (2018). Using hierarchical linear models to examine Approximate Number System acuity: The role of trial-level and participant-level characteristics. *Frontiers in Psychology*, *9*(NOV), 1–14. https://doi.org/10.3389/fpsyg.2018.02081

Brannon, E. M., Lutz, D., & Cordes, S. (2006). The development of area discrimination and its implications for number representation in infancy. *Developmental Science*, *9*(6), F59–F64. https://doi.org/10.1111/j.1467-7687.2006.00530.x

Butterworth, B. (2010). Foundational numerical capacities and the origins of dyscalculia. *Trends in Cognitive Sciences*, *14*(12), 534–541. https://doi.org/10.1016/j.tics.2010.09.007

Castaldi, E., Turi, M., Gassama, S., Piazza, M., & Eger, E. (2020). Excessive visual crowding effects in developmental dyscalculia. *Journal of Vision*, *20*, 1–20. https://doi.org/10.1101/2020.03.16.993972

Cavanagh, J. P. (1972). Relation between the immediate memory span and the memory search rate. *Psychological Review*, *79*(6), 525–530. https://doi.org/10.1037/h0033482

Cheng, D., Xiao, Q., Cui, J., Chen, C., Zeng, J., Chen, Q., & Zhou, X. (2020). Short-term numerosity training promotes symbolic arithmetic in children with developmental dyscalculia: The mediating role of visual form perception. *Developmental Science*, *23*(4), 1–8. https://doi.org/10.1111/desc.12910

Cicchini, G. M., Anobile, G., & Burr, D. C. (2016). Spontaneous perception of numerosity in humans. *Nature Communications*, *7*, 1–7. https://doi.org/10.1038/ncomms12536

Clarke, S., & Beck, J. (2021). The number sense represents (rational) numbers. *Behavioral and Brain Sciences*, *44*, e178. https://doi.org/10.1017/S0140525x21000571

Clayton, S., & Gilmore, C. (2015). Inhibition in dot comparison tasks. *ZDM - Mathematics Education*, *47*(5), 759–770. https://doi.org/10.1007/s11858-014-0655-2

Clayton, S., Gilmore, C., & Inglis, M. (2015). Dot comparison stimuli are not all alike: The effect of different visual controls on ANS measurement. *Acta Psychologica*, *161*, 177–184. https://doi.org/10.1016/j.actpsy.2015.09.007

Cordes, S., & Brannon, E. M. (2008). The difficulties of representing continuous extent in infancy: Using number is just easier. *Child Development*, *79*(2), 476–489. https://doi.org/10.1111/j.1467-8624.2007.01137.x

Dakin, S. C., Tibber, M. S., Greenwood, J. A., Kingdom, F. A. A., & Morgan, M. J. (2011). A common visual metric for approximate number and density. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(49), 19552–19557. https://doi.org/10.1073/pnas.1113195108

DeWind, N. K., Adams, G. K., Platt, M. L., & Brannon, E. M. (2015). Modeling the approximate number system to quantify the contribution of visual stimulus features. *Cognition*, *142*, 247–265. https://doi.org/10.1016/j.cognition.2015.05.016

DeWind, N. K., Bonner, M. F., & Brannon, E. M. (2020). Similarly oriented objects appear more numerous. *Journal of Vision*, *20*(4), 1–11. https://doi.org/10.1167/jov.20.4.4

DeWind, N. K., & Brannon, E. M. (2012). Malleability of the approximate number system: Effects of feedback and training. *Frontiers in Human Neuroscience*, *6*(APRIL), 1–10. https://doi.org/10.3389/fnhum.2012.00068

DeWind, N. K., & Brannon, E. M. (2016). Significant inter-test reliability across approximate number system assessments. *Frontiers in Psychology*, *7*(MAR), 1–10. https://doi.org/10.3389/fpsyg.2016.00310

Dowker, A., & Kaufmann, L. (2009). Atypical development of numerical cognition: Characteristics of developmental dyscalculia. *Cognitive Development*, *24*(4), 339–342. https://doi.org/10.1016/j.cogdev.2009.09.010

Droit-Volet, S., Clement, A., & Fayol, M. (2008). Time, number and length: Similarities and differences in discrimination in adults and children. *Quarterly Journal of Experimental Psychology*, *61*(12), 1827–1846. https://doi.org/10.1080/17470210701743643

Durgin, F. H. (1995). Texture density adaptation and the perceived numerosity and distribution of texture. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(1), 149–169. https://doi.org/10.1037/0096-1523.21.1.149

Durgin, F. H. (2008). Texture density adaptation and visual number revisited. *Current Biology*, *18*(18), 855–856. https://doi.org/10.1016/j.cub.2008.07.053

Elliott, L., Feigenson, L., Halberda, J., & Libertus, M. E. (2019). Bidirectional, longitudinal associations between math ability and approximate number system precision in childhood. *Journal of Cognition and Development*, *20*(1), 56–74. https://doi.org/10.1080/15248372.2018.1551218

Feigenson, L. (2007). The equality of quantity. *Trends in Cognitive Sciences*, *11*(5), 185–187. https://doi.org/10.1016/j.tics.2007.01.006

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*(7), 307–314. https://doi.org/10.1016/j.tics.2004.05.002

Ferrigno, S., Jara-Ettinger, J., Piantadosi, S. T., & Cantlon, J. F. (2017). Universal and uniquely human factors in spontaneous number perception. *Nature Communications*, *8*, 13968. https://doi.org/10.1038/ncomms13968

Franconeri, S. L., Bemis, D. K., & Alvarez, G. A. (2009). Number estimation relies on a set of segmented objects. *Cognition*, *113*(1), 1–13. https://doi.org/10.1016/j.cognition.2009.07.002

Fuhs, M. W., McNeil, N. M., Kelley, K., O'Rear, C., & Villano, M. (2016). The role of non-numerical stimulus features in approximate number system training in preschoolers from low-income homes. *Journal of Cognition and Development*, *17*(5), 737–764. https://doi.org/10.1080/15248372.2015.1105228

Gebuis, T., Cohen Kadosh, R., & Gevers, W. (2016). Sensory-integration system rather than approximate number system underlies numerosity processing: A critical review. *Acta Psychologica*, *171*, 17–35. https://doi.org/10.1016/j.actpsy.2016.09.003

Gebuis, T., & Reynvoet, B. (2011). Generating nonsymbolic number stimuli. *Behavior Research Methods*, *43*(4), 981–986. https://doi.org/10.3758/s13428-011-0097-5

Gilmore, C., Cragg, L., Hogan, G., & Inglis, M. (2016). Congruency effects in dot comparison tasks: Convex hull is more important than dot area. *Journal of Cognitive Psychology*, *28*(8), 923–931. https://doi.org/10.1080/20445911.2016.1221828

Halberda, J., & Feigenson, L. (2008). Developmental change in the acuity of the "number sense": The approximate number system in 3-, 4-, 5-, and 6-year-olds and adults. *Developmental Psychology*, *44*(5), 1457–1465. https://doi.org/10.1037/a0012682

Halberda, J., Ly, R., Wilmer, J. B., Naiman, D. Q., & Germine, L. (2012). Number sense across the lifespan as revealed by a massive Internet-based sample. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(28), 11116–11120. https://doi.org/10.1073/pnas.1200196109

Halberda, J., Mazzocco, M. M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, *455*(7213), 665–668. https://doi.org/10.1038/nature07246

Hannula, M. M., & Lehtinen, E. (2005). Spontaneous focusing on numerosity and mathematical skills of young children. *Learning and Instruction*, *15*(3), 237–256. https://doi.org/10.1016/j.learninstruc.2005.04.005

Hurewitz, F., Gelman, R., & Schnitzer, B. (2006). Sometimes area counts more than number. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(51), 19599–19604. https://doi.org/10.1073/pnas.0609485103

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259. https://doi.org/10.1109/34.730558

Izard, V., Sann, C., Spelke, E. S., & Streri, A. (2009). Newborn infants perceive abstract numbers. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(25), 10382–10385. https://doi.org/10.1073/pnas.0812142106

Jones, E., Oliphant, T., & Peterson, P. (2001). *SciPy: Open source scientific tools for Python.*

Kleinschmidt, A., Büchel, C., Zeki, S., & Frackowiak, R. S. J. (2002). Human brain activity during spontaneously reversing perception of ambiguous figures. *Biomedical Imaging V - Proceedings of the 5th IEEE EMBS International Summer School on Biomedical Imaging, SSBI 2002*, *September* (pp. 2427–2433). https://doi.org/10.1109/SSBI.2002.1233971

Leibovich, T., Katzin, N., Harel, M., & Henik, A. (2017). From "sense of number" to "sense of magnitude": The role of continuous magnitudes in numerical cognition. *Behavioral and Brain Sciences*, *40*, e164. https://doi.org/10.1017/S0140525x16000960

Libertus, M. E., & Brannon, E. M. (2010). Stable individual differences in number discrimination in infancy. *Developmental Science*, *13*(6), 900–906. https://doi.org/10.1111/j.1467-7687.2009.00948.x

Libertus, M. E., Feigenson, L., & Halberda, J. (2011). Preschool acuity of the approximate number system correlates with school math ability. *Developmental Science*, *14*(6), 1292–1300. https://doi.org/10.1111/j.1467-7687.2011.01080.x

Libertus, M. E., Feigenson, L., & Halberda, J. (2013). Is approximate number precision a stable predictor of math ability? *Learning and Individual Differences*, *25*, 126–133. https://doi.org/10.1016/j.lindif.2013.02.001

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

Mazzocco, M. M. M., Feigenson, L., & Halberda, J. (2011). Preschoolers' precision of the approximate number system predicts later school mathematics performance. *PLoS ONE*, *6*(9), e23749. https://doi.org/10.1371/journal.pone.0023749

Morgan, M. J., Raphael, S., Tibber, M. S., & Dakin, S. C. (2014). A texture-processing model of the "visual sense of number." *Proceedings of the Royal Society B: Biological Sciences*, *281*(1790), 1–9. https://doi.org/10.1098/rspb.2014.1137

Norris, J. E., Clayton, S., Gilmore, C., Inglis, M., & Castronovo, J. (2019). The measurement of approximate number system acuity across the lifespan is compromised by congruency effects. *Quarterly Journal of Experimental Psychology*, *72*(5), 1037–1046. https://doi.org/10.1177/1747021818779020

O'Hearn, K., & Luna, B. (2009). Mathematical skills in Williams syndrome: Insight into the importance of underlying representations. *Developmental Disabilities Research Reviews*, *15*(1), 11–20. https://doi.org/10.1002/ddrr.47

Odic, D. (2018). Children's intuitive sense of number develops independently of their perception of area, density, length, and time. *Developmental Science*, *21*(2), 1–15. https://doi.org/10.1111/desc.12533

Odic, D., Hock, H., & Halberda, J. (2014). Hysteresis affects approximate number discrimination in young children. *Journal of Experimental Psychology: General*, *143*(1), 255–265. https://doi.org/10.1037/a0030825. Hysteresis

Odic, D., Libertus, M. E., Feigenson, L., & Halberda, J. (2013). Developmental change in the acuity of approximate number and area representations. *Developmental Psychology*, *49*(6), 1103–1112. https://doi.org/10.1037/a0029472

Odic, D., Pietroski, P., Hunter, T., Lidz, J., & Halberda, J. (2013). Young children's understanding of "more" and discrimination of number and surface area. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 451–461. https://doi.org/10.1037/a0028874

Park, J. (2022). Flawed stimulus design in additive-area heuristic studies. *Cognition*, *229*(April), 104919. https://doi.org/10.1016/j.cognition.2021.104919

Piazza, M., De Feo, V., Panzeri, S., & Dehaene, S. (2018). Learning to focus on number. *Cognition*, *181*(July), 35–45. https://doi.org/10.1016/j.cognition.2018.07.011

Piazza, M., Facoetti, A., Trussardi, A. N., Berteletti, I., Conte, S., Lucangeli, D., Dehaene, S., & Zorzi, M. (2010). Developmental trajectory of number acuity reveals a severe impairment in developmental dyscalculia. *Cognition*, *116*(1), 33–41. https://doi.org/10.1016/j.cognition.2010.03.012

Piazza, M., Izard, V., Pinel, P., Le Bihan, D., & Dehaene, S. (2004). Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron*, *44*(3), 547–555. https://doi.org/10.1016/j.neuron.2004.10.014

Pica, P., Lemer, C., Izard, V., & Dehaene, S. (2004). Exact and approximate arithmetic in an Amazonian indigene group. *Science*, *306*(5695), 499–503. https://doi.org/10.1126/science.1102085

Picon, E., Dramkin, D., & Odic, D. (2019). Visual illusions help reveal the primitives of number perception. *Journal of Experimental Psychology: General*, *148*(10), 1675–1687. https://doi.org/10.1037/xge0000553

Price, G. R., Palmer, D., Battista, C., & Ansari, D. (2012). Nonsymbolic numerical magnitude comparison: Reliability and validity of different task variants and outcome measures, and their relationship to arithmetic achievement in adults. *Acta Psychologica*, *140*(1), 50–57. https://doi.org/10.1016/j.actpsy.2012.02.008

Purpura, D. J., & Simms, V. (2018). Approximate number system development in preschool: What factors predict change? *Cognitive Development*, *45*(February), 31–39. https://doi.org/10.1016/j.cogdev.2017.11.001

Smets, K., Sasanguie, D., Szűcs, D., & Reynvoet, B. (2015). The effect of different methods to construct non-symbolic stimuli in numerosity estimation and comparison. *Journal of Cognitive Psychology*, *27*(3), 310–325. https://doi.org/10.1080/20445911.2014.996568

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

Szűcs, D., & Myers, T. (2017). A critical analysis of design, facts, bias and inference in the approximate number system training literature: A systematic review. *Trends in Neuroscience and Education*, *6*(August), 187–203. https://doi.org/10.1016/j.tine.2016.11.002

Szűcs, D., Nobes, A., Devine, A., Gabriel, F. C., & Gebuis, T. (2013). Visual stimulus parameters seriously compromise the measurement of approximate number system acuity and comparative effects between adults and children. *Frontiers in Psychology*, *4*(JUL), 1–12. https://doi.org/10.3389/fpsyg.2013.00444

Tomlinson, R. C., DeWind, N. K., & Brannon, E. M. (2020). Number sense biases children's area judgments. *Cognition*, *204*(March), 104352. https://doi.org/10.1016/j.cognition.2020.104352

Treisman, A. (1982). Perceptual grouping and attention in visual search for features and for objects. *Journal of Experimental Psychology: Human Perception and Performance*, *8*(2), 194–214. https://doi.org/10.1037/0096-1523.8.2.194

Wang, J., Halberda, J., & Feigenson, L. (2017). Approximate number sense correlates with math performance in gifted adolescents. *Acta Psychologica*, *176*(April), 78–84. https://doi.org/10.1016/j.actpsy.2017.03.014

Wilkey, E. D., & Price, G. R. (2019). Attention to number: The convergence of numerical magnitude processing, attention, and mathematics in the inferior frontal gyrus. *Human Brain Mapping*, *40*(3), 928–943. https://doi.org/10.1002/hbm.24422

Wilson, A. J., Revkin, S. K., Cohen, D., Cohen, L., & Dehaene, S. (2006). An open trial assessment of "the number race", an adaptive computer game for remediation of dyscalculia. *Behavioral and Brain Functions*, *2*, 1–16. https://doi.org/10.1186/1744-9081-2-20

Yousif, S. R., & Keil, F. C. (2019). The additive-area heuristic: An efficient but illusory means of visual area approximation. *Psychological Science*, *30*(4), 495–503. https://doi.org/10.1177/0956797619831617

## Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supplementary Materials A: Model validation & Supplementary Materials B: Reaction time analysis
Table S1
Table S2
Table S3
Table S4
Table S5