Cognition 213 (2021) 104805

Contents lists available at ScienceDirect

Cognition

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

Bias and noise in proportion estimation: A mixture psychophysical model☆

Camilo Gouet [a,c,*], Wei Jin [b], Daniel Q. Naiman [b], Marcela Peña [c], Justin Halberda [a]

[a] Department of Psychological and Brain Sciences, The Johns Hopkins University, Baltimore, MD 21218, USA
[b] Department of Applied Mathematics and Statistics, The Johns Hopkins University, Baltimore, MD 21218, USA
[c] Laboratorio de Neurociencias Cognitivas, Escuela de Psicología, Pontificia Universidad Católica de Chile, Avenida Vicuña Mackenna 4860, Macul, Santiago, Chile

ARTICLE INFO

Keywords:
Psychophysics
Proportion estimation
Bias
Noise
Mental magnitudes

ABSTRACT

The importance of proportional reasoning has long been recognized by psychologists and educators, yet we still do not have a good understanding of how humans mentally represent proportions. In this paper we present a psychophysical model of proportion estimation, extending previous approaches. We assumed that proportion representations are formed by representing each magnitude of a proportion stimuli (the part and its complement) as Gaussian activations in the mind, which are then mentally combined in the form of a proportion. We next derived the internal representation of proportions, including bias and internal noise parameters -capturing respectively how our estimations depart from true values and how variable estimations are. Methodologically, we introduced a mixture of components to account for contaminating behaviors (guessing and reversal of responses) and framed the model in a hierarchical way. We found empirical support for the model by testing a group of 4th grade children in a spatial proportion estimation task. In particular, the internal density reproduced the asymmetries (skewedness) seen in this and in previous reports of estimation tasks, and the model accurately described wide variations between subjects in behavior. Bias estimates were in general smaller than by using previous approaches, due to the model's capacity to absorb contaminating behaviors. This property of the model can be of especial relevance for studies aimed at linking psychophysical measures with broader cognitive abilities. We also recovered higher levels of noise than those reported in discrimination of spatial magnitudes and discuss possible explanations for it. We conclude by illustrating a concrete application of our model to study the effects of scaling in proportional reasoning, highlighting the value of quantitative models in this field of research.

1. Introduction

Throughout history, inquiry in science has been aided when formal models are developed for observed phenomena. In cognitive science, these models often attempt to make sense of behavior while estimating the underlying parameters that guide and inspire it. In this context, psychophysical models have played a key role in our understanding of how we perceive and mentally combine magnitudes from the world, providing estimates to explain the biases and variability inherent in much of our perception. Yet most psychophysical models have accounted for only a portion of magnitude-related phenomena, namely, absolute magnitudes, as when we estimate the size of an object or the duration of an event.

While absolute magnitudes are important, magnitudes can also be combined to form ratios or proportions, and in many situations it makes more sense to focus on the relations of magnitudes rather than on them in isolation. When we use a map for example, we pay attention to the relative distances and positions in one space to be used in another space. That is, "twice as far" on the map (a distance of e.g., 2 in.) corresponds to "twice as far" in the world (a distance that can be arbitrarily long depending on the map scale). The importance of proportions has long been recognized by psychologists (Inhelder & Piaget, 1958; Siegler & Vago, 1978), and impacts a wide variety of domains including the perception of shapes (Sophian, 2000), probability judgment (Acredolo, Connor, Banks, & Horobin, 1989), music perception (Plantinga & Trainor, 2005), and mathematics (Lesh, Post, & Behr, 1988; Siegler, Fazio, Bailey, & Zhou, 2013), among others. Yet, surprisingly, we have fewer models, which are less developed, to describe how we mentally

☆ This paper is a part of special issue "Special Issue in Honour of Jacques Mehler, Cognition's founding editor".

* Corresponding author at: Laboratorio de Neurociencias Cognitivas, Escuela de Psicología, Pontificia Universidad Católica de Chile, Avenida Vicuña Mackenna 4860, Macul, Santiago, Chile

E-mail addresses: camilogouet@gmail.com (C. Gouet), wjin@jhu.edu (W. Jin), daniel.naiman@jhu.edu (D.Q. Naiman), mpenag@uc.cl (M. Peña), halberda@jhu.edu (J. Halberda).

https://doi.org/10.1016/j.cognition.2021.104805
Received 15 September 2020; Received in revised form 4 June 2021; Accepted 5 June 2021
Available online 23 June 2021

represent proportions than those we have for absolute magnitudes (Chen & Verguts, 2017; Jacob, Vallentin, & Nieder, 2012; Matthews, Lewis, & Hubbard, 2016; Shepard, 1981). Are the representations of absolute magnitudes different from those of relative magnitudes? Can we use what we know from the former to understand and model the latter? Are the insights already seen in modeling absolute magnitudes also expected for the modeling of relative magnitudes?

In this study we contributed to efforts in addressing these and other questions with a model of proportion representation. Our approach was to use the Cyclical Power Model by Hollands and Dyre (2000) as a starting point. This model advanced the idea that proportion representations result from the mental combination of absolute magnitude representations. While promising and widely used, this model does not incorporate any effects of internal noise and relies only on bias to fit subject/group variance. We modified this model to include parameters for both bias and internal noise and tested its performance in behavioral data. We found these modifications to be central to providing an accurate estimate of subject performance.

Why are such modifications necessary? Many experiments have shown that our judgment of proportions is not accurate, but rather, admits of striking biases. In a typical experiment, subjects are presented with two stimuli (e.g. two tones of different durations; two shapes of different sizes) and the task is to estimate the magnitude of one stimulus relative to the whole magnitude (part/whole judgment). By plotting estimates against true proportion values, we can easily detect biases as departures from the diagonal line (perfect performance). Hollands and Dyre (2000) noted that for some continua, these distortions away from the diagonal follow a pattern of overestimation for true proportions less than 0.5 and underestimation above 0.5, while the reverse also happens (underestimation followed by overestimation), giving rise to an s-shaped curve illustrated in Fig. 1. These behaviors have been observed for many continua including line lengths (Spence, 1990), graphical elements (Spence & Krizel, 1994), spatial position (Huttenlocher, Hedges, & Duncan, 1991; Zax et al., 2019), numerosities (Varey, Mellers, & Birnbaum, 1990; H. Zhang & Maloney, 2012) and time intervals (Nakajima, 1987) among others (reviewed in Holland & Dyre, 2000). Different theories have been proposed to explain these patterns (Huttenlocher et al., 1991; Nakajima, 1987; Spence & Krizel, 1994), but a very promising one is the Cyclical Power Model. According to this model (developed after Spence, 1990), these biases arise from the biases in representing each of the magnitudes composing the stimuli (the part and the whole), which are described by power functions of the physical magnitudes (Steven's power law) (Stevens, 1957). The mental combination of such biased representations in a proportion is what gives rise to the s-shaped curves shown in Fig. 1A.

This 'one-cycle' version of CPM, in Fig. 1A, was then generalized to account for multi-cycle biases, e.g., in Fig. 1B. Indeed, studies using graphical shapes and geometrical regions (Huttenlocher et al., 1991; Spence & Krizel, 1994), showed response patterns consisting of compressed repetitions of the one-cycle behavior, suggesting that subjects can mentally split the stimulus and/or the response space and use intermediate reference points, thus increasing their accuracy overall. The simplicity and explanatory power of CPM is noteworthy, and it is currently used as a tool for measuring proportional reasoning in different tasks including the popular number line task (Barth & Paladino, 2011; Slusser, Santiago, & Barth, 2013; Zax, Slusser, & Barth, 2019), and spatial position reproduction tasks (Zax, Williams, et al., 2019). But besides these merits CPM offers only a partial picture of proportion representation.

Missing from this model are estimates of internal noise, which together with bias form two key elements to characterize mental magnitudes. According to behavioral and neural evidence, from the early sensory processing of incoming stimuli up to the execution of a response, the internal signal passes through a series of processing steps that may corrupt it in random ways, due in part to the intrinsic variability of brain dynamics (Garrett, Kovacevic, Mcintosh, & Grady, 2011) and neural activity more generally (Pardo-vazquez et al., 2019). This internal 'noise' infuses a stochastic character to mental magnitudes, causing variability in our estimations, and by studying systematic patterns of variability in the responses it is possible to infer the noise associated with the perception of a given stimulus (see review in (Odic, Im, Eisinger, Ly, & Halberda, 2016; Wei & Stocker, 2015).[1] The importance of measuring this parameter has been clearly demonstrated in studies with absolute magnitudes. For example, Odic, Libertus, Feigenson, and Halberda (2013) obtained measures of internal noise for the perception of area and approximate numerosity in different age groups, revealing great variations between subjects that helped disentangle developmental trajectories which were obscured when using average measures of accuracy. Similarly, estimates of noise have been used to study connections between cognitive systems (e.g. Lourenco & Bonny, 2016), and have guided the search for neural tuning curves in the brain (Nieder & Dehaene, 2009). These advances were made possible by taking internal noise seriously and attempting to model it, rather than failing to engage with it, as is generally current practice for proportion representation (e.g. Boyer & Levine, 2012; Boyer, Levine, & Huttenlocher, 2008; Vukovic et al., 2014).

Furthermore, this situation for absolute magnitudes contrasts sharply with our knowledge of internal noise for the perception of relative magnitudes and proportions. A handful of brain studies have shown that the representation of relative magnitudes is also characterized by 'noisy' activations in similar fronto-parietal regions as those seen for the representation of absolute magnitudes (e.g. Jacob et al., 2012; Lewis, Matthews, & Hubbard, 2015). However, the field has yet to solve other puzzles related to noise. Earlier studies (that gave rise to the formulation of CPM) noted that proportion estimations do not distribute symmetrically around central tendency values (i.e. are skewed) (Spence, 1990; Spence & Krizel, 1994. Spence and Krizel (1994) proposed that the skewedness in proportion estimation distribution arise from the action of repulsive forces that push estimations away from the boundaries. These forces were thought to result from the use of imperfect strategies on the part of subjects, which leave no room for responses to fall outside of the boundaries while leaving ample space for responses to be distributed away from the boundaries resulting in skewed or asymmetric distributions. This distributional aspect was used to explain the origin of biases but no attempt to model it nor to derive an internal noise estimate was achieved. This lack of focus on internal noise can also be seen in models of probability/frequency judgments, where s-shaped distortions have been well documented, but where the focus has been in understanding the causes of these biases (see review in H. Zhang & Maloney, 2012). More recently, Rouder & Geary (2014) noted this gap and modeled intrasubject variability in proportion estimation. In particular, they pointed out 1) the implausibility of assuming equal normal errors in proportion estimation, as is typically done when fitting CPM both in earlier but also in more recent studies (e.g. Barth & Paladino, 2011) and 2) the problems of using central tendency measures (i.e. aggregated group means) that do not capitalize on within-subject variability. They thus modified CPM and introduced a random error term that predicted varying levels of variability with presented proportions. This approach allowed them to describe developmental trajectories of performance in the number line task with more detail than by using CPM in the traditional way (i.e. traditional use of central tendency measures).

This improved version of CPM highlighted the value of modeling variability of responses in proportion estimation, but it focused more on 'curve fitting' than on investigating the cognitive sources of such variability (see Lee, 2018 for a discussion of approaches in cognitive

[1] As noted by (Lockhead, 2004), the term 'noise' may be incorrect because variability is actually structured, and Halberda (2016) and Halberda and Odic (2015) stress that 'confidence' may be the more appropriate term, but it has been used here for historical reasons derived from its use for the presentation of Signal Detection Theory (Green & Sweet, 1966).

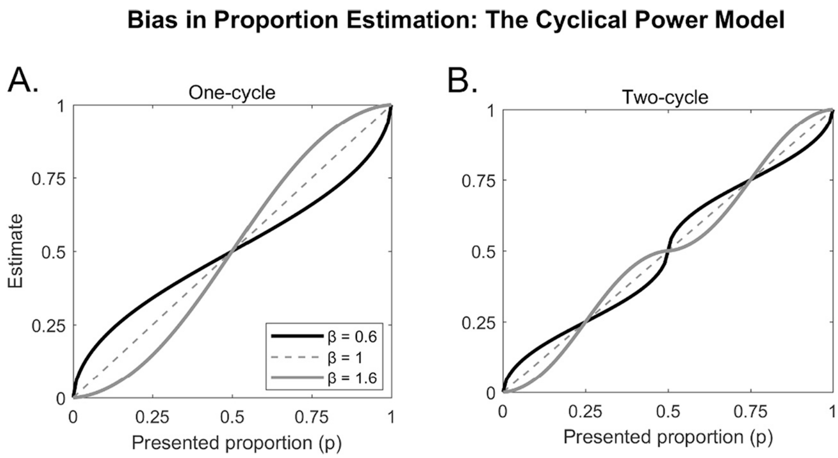

Fig. 1. Bias in proportion estimation and the Cyclical Power Model (CPM). The curves illustrate the different types of biases in proportion estimation. They are plots of CPM's equation ($y = p^{\beta}/(p^{\beta} + (1\text{- } p)^{\beta})$), which has one free parameter β. **A**, when $\beta < 1$, responses are over- and then underestimated and when $\beta > 1$ the reverse is observed; no bias is seen when $\beta = 1$. **B**, the two-cycle CPM predicts the same pattern as in A, but scaled to the middle and repeated two times. It arises from the use of the half as a reference point, in addition to the references from the lower and upper ends (0,1).

modeling). Indeed, Rouder & Geary's (2014) purpose was to accurately identify which variant of CPM (one-cycle, two-cycle or other candidate models) better accounted for children's behavior in the number line task, and in so doing, variability was described but not related to its potential cognitive sources. Our study was thus motivated to fill this gap in knowledge, focusing on studying bias and response variability in proportion estimation from a more psychological perspective. Ours and Rouder & Geary's (2014) are complimentary approaches.

Our modeling approach is schematized in Fig. 2, along with the task we implemented. We followed the logic of cognitive generative models (e.g. Kemp, Bernstein, & Tenenbaum, 2005; for an overview see Lee, 2011), in the sense of conceiving complex representations (i.e. proportion) as the combination of simpler ones (i.e. absolute magnitudes). This was the rationale of CPM, and here we pushed it further and derived the internal representation of a proportion by fully accounting for the statistical properties of absolute magnitude representations.

As in previous modeling approaches (e.g. Odic et al., 2016; Wei & Stocker, 2015), we assumed a Gaussian noise structure for absolute magnitudes representation, that fed into the proportion representations. Specifically, we assumed that physical magnitudes are mentally represented as Gaussian distributions ordered along an internal scale with two parameters (β and w): β controls the position of the Gaussians on the scale and thus determines the degree of bias, and w controls the variability of the distributions along the internal scale, which we assume to increase in proportion to their mean (scalar variability). The parameter w is typically called the Weber Fraction and it is an index of internal noise. Although not without criticism (e.g. Inglis & Gilmore, 2014), ample behavioral and neural evidence gives support to this account (e.g. Grondin & Killeen, 2009; Halberda & Odic, 2015; Pardo-vazquez et al., 2019; Piazza, Izard, Pinel, Le Bihan, & Dehaene, 2004; Whalen, Gallistel, & Gelman, 1999), as does the potential success of our current modeling effort.

To fit our model, we gathered empirical proportion estimation response data from a group of 4th grade children, tested in the proportion estimation game of Fig. 2 (Gouet, Carvajal, Halberda, & Peña, 2020; Möhring, Newcombe, Levine, & Frick, 2015; Spence & Krizel, 1994). Different variants of this task have been implemented (Boyer et al., 2008; Boyer & Levine, 2015) and studies show that children of this age (9–10 years old) can readily solve proportional reasoning tasks while typically exhibiting wide variations between subjects and larger errors and biases compared to adult participants (Möhring et al., 2015), thus offering ideal opportunities to characterize mental processes at work (Möhring, Frick, & Newcombe, 2018; Spence & Krizel, 1994).

While somewhat atypical, we would like to preview some of our results here in order to motivate readers by seeing the problems of noise and bias within the raw data itself. The total of all subject responses from the proportion estimation task from this article are shown in Fig. 3A. We can see that the bulk of responses fall along the main diagonal, as subjects tracked the presented proportions. However, further inspection also suggests two additional groups of responses: one located along the anti-diagonal, suggesting a reversal of responses, and another group scattering more homogenously across the response space, suggesting random guesses. These behaviors were also seen at the individual level (Fig. 3C), replicating previous observations (e.g. Spence, 1990), and

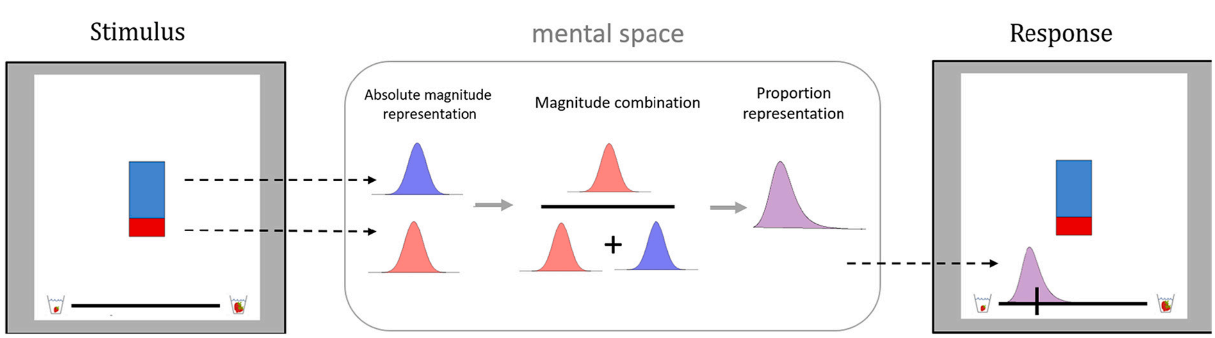

Fig. 2. Schematics of the proposed model and the task implemented. In each trial of the task, subjects see two-colored columns and judge the proportion of red size to the total size by clicking on the response bar. The task is framed as a juice game, and the cartoons on the left and right sides serve as anchor points, indicating a weak and strong flavor of juice respectively (see instructions in methods). In this example $p = 0.25$. According to our model, the red and blue sizes are mentally represented as Gaussian activations with some bias and internal noise (red and blue humps). These representations are then mentally combined, generating an internal representation of a proportion, which is schematically mapped onto the response line. Responses are then modeled as samples from this internal representation, and by analyzing the whole pattern of responses we can recover the level of bias and noise in the original Gaussians. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

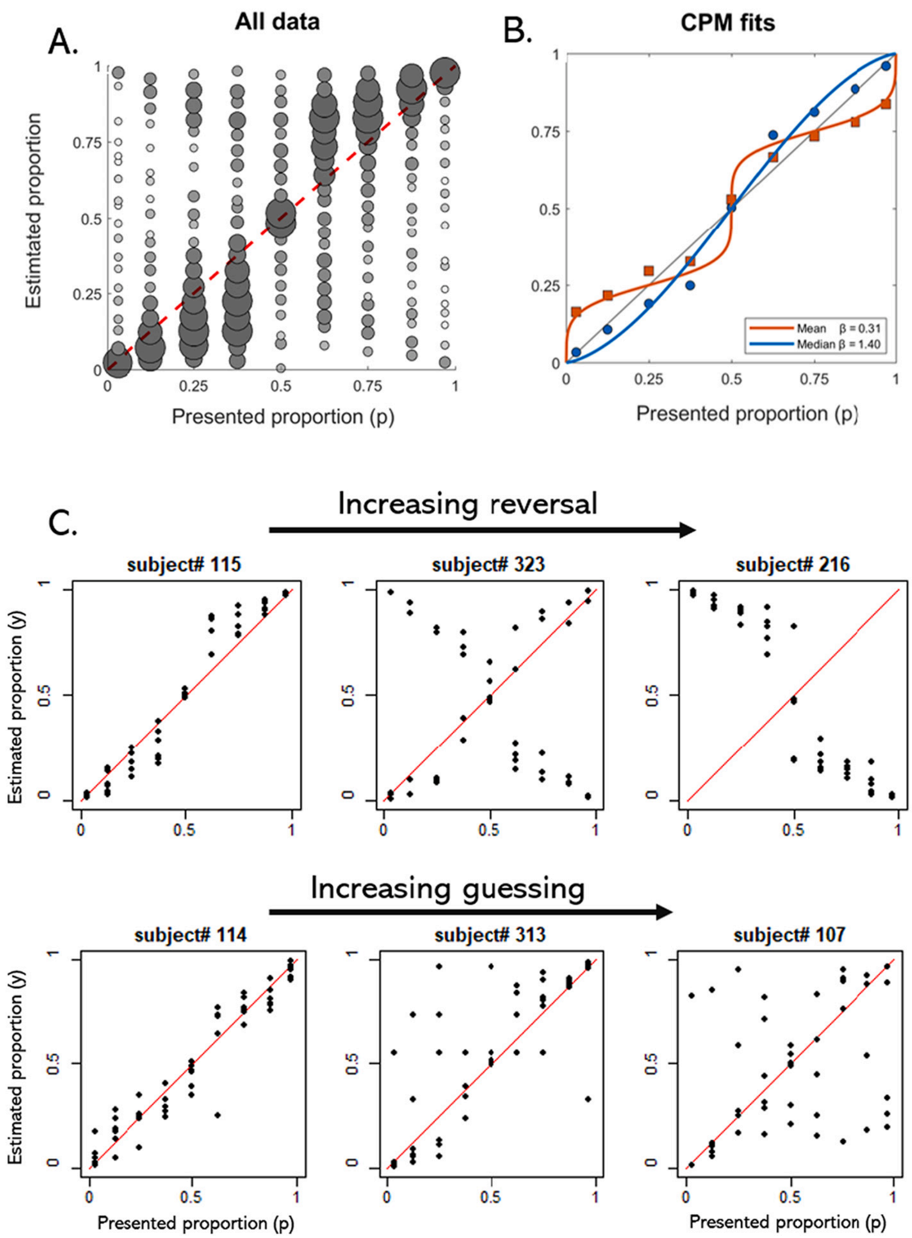

Fig. 3. Estimation responses and CPM traditional approach. **A**, bubble plot of the full data set. The radius and color intensity of each dot is proportional to the number of responses falling near each point (20 bins). **B**, mean and median estimates fitted to CPM using least squares. The curves show the model that best explain each measure (based on AICc values). Two-cycle model for the means, one-cycle for the medians. β estimates are in the legend. **C**, sample of individual data, with varying levels of reversal and guessing behaviors.

together they formed the building blocks of our mixture model.

Traditionally, estimation data are fit to CPM using central tendency measures (means or medians) and least squares procedures (Barth & Paladino, 2011; Hollands & Dyre, 2000; Slusser et al., 2013; Spence & Krizel, 1994). This approach is shown at the group level in Fig. 3B (earlier work by Spence & Krizel, 1994; Hollands & Dyre, 2000, used aggregated means to fit models; most recent work from for example Barth & colleagues include both group and subject level measures). Although for both measures we observe s-shaped biases well described by CPM curves (consistent with previous results, Spence & Krizel, 1994; Hollands & Dyre, 2000), there are major differences between them, with means being best explained by a two-cycle model with $\beta_{mean} = 0.31$, and medians by a one-cycle model with $\beta_{median} = 1.4$). This implies qualitative differences between the two measures in terms of model selected and parameter estimates and reveals a problem of inconsistent fits to the data. The heavy tails and skewness in the distribution of responses likely explain these differences and render this approach unsuitable in this context.

Similar concerns were raised by Rouder & Geary (2014), which motivated their modeling efforts to better account for within-subject variance. Yet Rouder & Geary's model did not include contaminating behaviors (Fig. 3C) which can distort parameter estimates and their psychological interpretation. These methodological considerations, along with the idea of building a more principled model grounded in our knowledge of absolute magnitudes, motivated our proposal.

As mentioned, the presence of contaminating responses motivated an extension of the original formulation having only an internal component. Specifically, we observed responses consistent with random guessing, possibly due to inattention or boredom, and we also observed the reversal of responses. Reversals, as shown above, appear when a participant presented with a proportion of say, 0.25 gave a response around 0.75. These behaviors were also reported by Spence (1990), examining adult participants in proportion estimation tasks with graphical displays. Contaminating behaviors have been observed in other tasks, including reaction time experiments (Ratcliff & Tuerlinckx, 2002), working memory (W. Zhang & Luck, 2008), discrimination task (Pica, Lemer, Izard, & Dehaene, 2004) and strategy-choice studies (Lee, Gluck, & Walsh, 2016). Misclassifying these behaviors as signal can seriously affect parameter estimates and their psychological interpretation, and researchers have advocated for methods to deal with them in formal ways (for review of different aproaches see Lee, 2018; Zeigenfuse & Lee, 2010). Thus, we complemented the internal component of the model described above with two contaminating pieces, representing guessing and reversal contributions. This extension resulted in a mixture model, that we framed in a hierarchical (multilevel) way.

We had three main goals: 1) to derive an internal representation that mimics the asymmetry (skewness) seen in the distribution of proportion representations (Spence, 1990; Spence & Krizel, 1994); 2) to determine whether bias and internal noise affect proportion estimation in a manner consistent with that reported for the judgment of absolute areas or lengths (because of our generative framework); and 3) to determine if these additions to the model will result in smaller biases (approaching 1) when fitting, compared to fitting CPM in the traditional way, due to our model's capacity to account for contaminating behaviors.

Finally, we aimed at illustrating a concrete application of the model for the study of scaling effects in proportional reasoning. The scaling factor in a proportional reasoning task quantifies the degree of disparity between two spaces that are put in correspondence, like the scales used in maps. Recent studies have systematically studied the effects of this factor, showing that subjects produce more errors of estimation as the scale increases, and that children are more sensitive to this element than adults (Boyer & Levine, 2012; Möhring et al., 2015; Vasilyeva & Huttenlocher, 2004). These effects are typically studied using overall measures of error, that may hide subtleties of both behavior and representation. Here we manipulated the ratio between the total height of the columns and the width of the response line in a systematic way and showed how modeling bias and noise can bring further insight to this phenomenon.

2. Materials and methods

2.1. Participants

We evaluated fifty seven, healthy and typically developing 4th grade children (29 females, mean age, 9.4 years, range: 9–11 years) who were recruited from a public school in Santiago, Chile. This study was part of a larger intervention aimed at training math abilities in children and the data presented here were taken before the intervention began. The children and their parents/caregivers gave written assent and consent, respectively, before participating in the study. The study received the corresponding ethics board approval before beginning.

2.2. Task and procedure

The proportion estimation task was modeled after Möhring et al. (2015) (see also Spence & Krizel, 1994). During the task, children saw a two-colored column appear in the center of a computer screen, with the lower section colored red to represent strawberry juice and the upper section colored blue to represent water (Fig. 2). The task was to indicate how much the mixture of juice and water would taste of strawberry after mixing it up, by clicking on a rating scale located below the column. The appearance of each stimulus was controlled by subjects, by pressing the space bar on the keyboard, and was present until the subject responded or until a 5000 ms timeout.

The set of stimuli was composed of 9 different proportions of juice to a total amount (0.03, 0.125, 0.25, 0.375, 0.5, 0.625, 0.75, 0.875, 0.97). These proportions were combined with three scaling factors (the ratio between the total height of the central column and the width of the response scale): 1:1, 1:2, and 1:3. Notice that this scaling factor ensured that children could not simply mentally rotate the central column and align it in order to generate an accurate answer on 1:2 and 1:3 trials. Rather, they must reason proportionally to generate their answer (as was done in other studies Möhring et al., 2018). Using this method, 27 different stimuli were created, each presented twice, in two blocks, totaling 54 stimuli. The order of presentation within each block was pseudorandom, avoiding more than 2 consecutive repetitions of the same scaling factor.

Children were evaluated in groups of 10 at their school, during class time, in a specially adapted room. There were 5 research assistants that evaluated two subjects each. Each participant was personally given the instructions of the task: "During the game you will see a column with a red part at the bottom, representing strawberry juice, very sweet, and a blue part above it representing water. Your task will be to indicate in this line below, how much the mixture of juice and water will taste of strawberry". They were also told that after some time the mixtures would disappear so they had to respond quickly but as accurate as possible.

There were 6 practice trials with the proportions: 0.95, 0.05, 0.7, 0.3, 0.45, 0.6. (scaling factor 1:1 for the first three, and scaling 1:2 for the last three). The first two trials served to indicate the right and left anchor points respectively, while the others helped children to understand they could use the response bar continuously (see Gouet et al., 2020; Möhring et al., 2015). Visual feedback was given in practice trials, in the form of a colored rectangle surrounding the perfect response in the response bar. A neutral beep was also played after each click. After completing the practice, children were asked if they had understood the game; if not, they were given the instructions and practice trials again. Only a few children required a second instruction run. Test trials followed immediately after the practice. Children received no feedback during test trials. For each trial we recorded the position in the rating scale and we included only test trials in the analysis.

2.3. Modeling

In this section we describe the structure of the models engaged for this study, placing more focus on mathematical aspects than on substantive (psychological) ones. The latter will be covered more thoroughly in the Results section.

2.3.1. The cyclical power model (CPM)

As mentioned in Introduction, the basic generative idea of Hollands and Dyre (2000)'s CPM is that the bias in proportion estimation results from the bias in mentally representing each magnitude composing the proportion stimulus. The equations of CPM are:

One-cycle:

$$Y_p = \frac{p^\beta}{p^\beta + (1-p)^\beta}$$

Two-cycle:

$$Y_p = \begin{cases} \dfrac{p^\beta}{p^\beta + (0.5-p)^\beta} * 0.5 & \text{if } p < 0.5 \\ \dfrac{(p-0.5)^\beta}{(p-0.5)^\beta + (1-p)^\beta} * 0.5 + 0.5 & \text{if } p => 0.5 \end{cases}$$

where Y_p corresponds to the mean or median of responses for each presented proportion p. Consistent with the literature (e.g. Barth & Paladino, 2011), we used a nonlinear least square method to fit CPM (one-cycle and two-cycle), to both the means and medians of the original data. Goodness of fit were evaluated with Akaike Information Criteria corrected for small samples (AICc) (Burnham & Anderson, 2002), as we fitted mean and median of responses.

2.3.2. Proportion estimation model (PEM)

As indicated in the Introduction, we built a mixture model with two main pieces: one related to the internal representation of proportions, and another related with contaminating behaviors. The internal component was more theoretically motivated, and we will describe it first. Together they formed our mixture model, that we framed as a hierarchical Bayesian mixture model (for an application of hierarchical Bayesian mixture models to the study of cognitive processes see Lee & Sarnecka, 2009); for an overview of these models see Lee (2018)). Some descriptions will be presented here and in Results, to emphasize or re-explain some key aspects of the model.

2.3.2.1. PEM internal representation function. The internal representation was derived by assuming that subjects represented each magnitude composing the stimuli (the sizes of red and blue columns) as Gaussian distributions, that were then mentally combined in the form of a proportion (Fig. 2). For each Gaussian, we assumed two properties: (1) they were centered in the magnitudes of the stimuli, compressed or expanded by a certain amount, and (2) they exhibited scalar variability (their standard deviations increased with their means). Thus we have:

$$INT_p = \frac{R}{R+B},$$

where INT is the random variable of the proportion representation; $R \sim N(p^\beta, w * p^\beta)$ and $B \sim N((1-p)^\beta, w * (1-p)^\beta)$ correspond to the red and blue magnitude representations respectively.[2] Note the similarity between this formulation and the one-cycle CPM, although here we use random variables. The parameter β controls the position of the Gaussians (or the degree of compression/expansion of the internal scale), and thus the amount of bias, and *w* is the Weber fraction, which controls the spread of the distributions (internal noise).

To find the probability density function (pdf) of INT (int_p), we used the Jacobian method, which follows from a change of variables in multivariate functions (see Rice (2007) and Supplemental Material for details of this derivation). We obtained the following expression:

$$int(z)_p = \frac{1}{tr^*(1-z)^2}\int_{-\infty}^{\infty} f_R\left(\frac{bz}{1-z}\right) f_B(b)|b|db$$

where $f_R(\cdot)$ and $f_B(\cdot)$ are the normal densities corresponding to R and B as defined above; *tr* is a truncation element (the difference in the cumulative distribution of INT between 1 and 0), which guarantees that the support for int_p is the interval [0,1]. In SM we provide an equivalent expression for $int(z)_p$, which uses the normal cumulative distribution, speeding up the computations when fitting the model.

2.3.2.1.1. Two-cycles. The internal component for the two-cycle model has the following equations:

$$INT_p = \begin{cases} \left(\frac{R}{R+B}\right)^* 0.5 & \text{if } p < 0.5 \\ \left(\frac{R}{R+B}\right)^* 0.5 + 0.5 & \text{if } p \geq 0.5 \end{cases}$$

When $p < 0.5$
$R \sim N(p^\beta, w^* p^\beta)$ and $B \sim N((0.5-p)^\beta, w^*(0.5-p)^{\beta\cdot})$
and when $p \geq 0.5$
$R \sim N((p-0.5)^\beta, w^*(p-0.5)^\beta)$ and $B \sim N((1-p)^\beta, w^*(1-p)^\beta)$.

The pdf of INT_p was derived using the same method as for the one-cycle model. Here we omit the formula, though there is very similar to that of the one-cycle model (see Supplementary Material).

2.3.2.2. Including contaminating elements. As indicated above, visual inspection of the data suggested the presence of two contaminating behaviors (in addition to the responses along the main diagonal that would be captured by the internal component). These contaminations were reversal of responses (located along the anti-diagonal) and responses distributed more homogenously, suggesting random guesses, perhaps due to inattention (Fig. 3A). We included these two pieces in a mixture model (see Zeigenfuse & Lee, 2010).

Thus, each response for a presented proportion *p* and for each participant *j*, $Y_{j,p}$, was modeled as a sample from the mixture distribution:

$$Y_{j,p} \sim \lambda_{int\ j} \boldsymbol{INT}_p(\beta_j, w_j) + \lambda_{rev\ j} \boldsymbol{REV}_p(\beta_j, w_j) + \lambda_{guess\ j} \boldsymbol{Guess}_p$$

where INT_p, Rev_p and $Guess_p$ are the internal and contaminating distributions with their respective weights ($\lambda_{int}, \lambda_{rev}\ \lambda_{guess}$ add up to 1). The reversal distribution was obtained by swapping *p* with *1-p* in the expression for the internal; for the guessing behavior we assumed a uniform distribution between 0 and 1 – but this function can be adjusted to fit any described pattern (e.g., guessing in the middle, avoiding end points, etc.). Determining a strategic guessing function for any particular task would be important future work.

2.4. Fitting PEM: A hierarchical Bayesian framework for PEM

We framed our model as a hierarchical Bayesian model (Fig. 4). We adopted this approach, as previous studies have found less variance in recovered parameters than when using maximum likelihood in hierarchical models, especially in cases with small sample sizes (Farrel & Ludwig, 2008; Wiecki, Sofer, & Frank, 2013). Under this approach, uncertainty is assumed on the part of parameters in the form of priors, and the focus is placed on deriving the full posterior distribution of parameters given the observed data (see reviews and textbooks Gelman et al., 2014; Hoff, 2009; Lee, 2018). A graphical depiction of our model is shown in Fig. 4.

In framing the model we used the indicator variable $Z_{i,j}$, as it facilitates computations in mixture models (see for example Zeigenfuse & Lee, 2010; Gelman et al., 2014). We assumed this variable to be generated from a categorical distribution parameterized by $\lambda_i = (\lambda_{i,\ int}, \lambda_{i,\ rev}, \lambda_{i,\ guess})$, with $i = 1, 2, \ldots, m$, subjects and $j = 1, 2, \ldots, n$ trials. Specifically, $Z_{i,j} = 1, 2$, and 3 indicates that the response for subject *i* at trial *j* is from internal, reversal, and random guessing, respectively. As the parameter of the categorical distribution, each component of λ_i represents the probability of each type of response for subject *i* in all the trials. Therefore, they must add up to one, i.e., $\lambda_{i,\ int} + \lambda_{i,\ rev} + \lambda_{i,\ guess} = 1$. The subject-level parameter λ_i follows from a Dirichlet distribution parametrized by $K = 3$ and group-level parameter $\alpha = (\alpha_1, \alpha_2, \alpha_3)$. Each component α_j when normalized to the sum of all the other components represents the group-level parameter for each λ_i (so for example, the group-level parameter of the reversal $\lambda_{i,\ rev}$ is $\alpha_2/(\alpha_1 + \alpha_2 + \alpha_3)$, and for guessing $\lambda_{i,\ guess}$ would be $\alpha_3/(\alpha_1 + \alpha_2 + \alpha_3)$. In Results we refer to them as *alpha_rev* and *alpha_guess* respectively.

In addition, conditional on the presented proportion $\boldsymbol{P}_{i,j}$, the response proportion $\boldsymbol{Y}_{i,\ j}$ is distributed from a mixture of internal, reversal, and uniform distributions, where the former two mixture components are parametrized by the parameter β_i and w_i.The subject-level parameters β_i and w_i follow from Gaussian distributions with group-level mean and variance parameters $(\mu_\beta, \sigma_\beta^2)$ and (μ_w, σ_w^2) (see Wiecki et al., 2013).

From the joint conditional posterior, we can derive marginal posterior distributions (full conditional posteriors) for each parameter of the model (see Hoff, 2009 for a pedagogical introduction to Bayesian analysis). Then we can obtain estimates of our interest (e.g. mean, medians, variance and credible intervals) and make statistical inferences.

The joint posterior distribution of all parameters was approximated using a Metropolis-Hasting algorithm (Hoff, 2009). We chose conjugate priors for the indicator variables and the probabilities for different type of responses (categorical and Dirichlet distributions respectively), allowing direct calculation of the posterior distribution for these parameters. The values of the hyperparameters were chosen to produce

[2] We use *p* and *(1- p)* to represent the red and blue magnitudes as this notation facilitates working with the equations in terms of *p* for both magnitudes. The value of p is obtained by dividing the red size by the whole and *(1-p)* by dividing the blue size by the whole. Equal results would be obtained if we used specific notations for the red size and blue size in the formulation of the Gaussians.

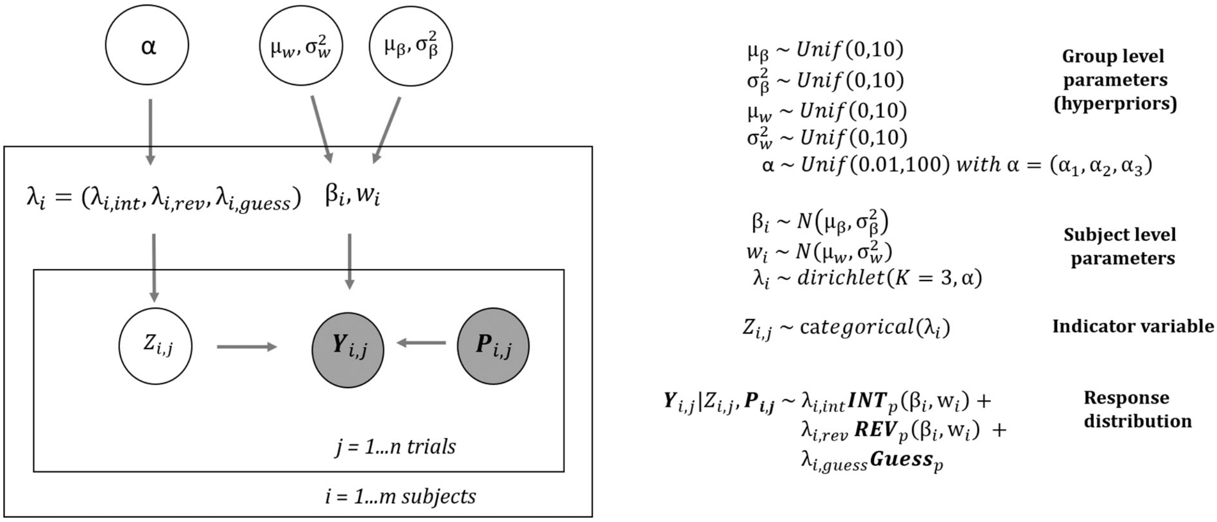

Fig. 4. Graphical representation of PEM. Each plate shows variables at the same level and arrows indicate dependencies between them. According to the hierarchical structure, each subject has their own set of parameters, that determine the response distribution. These parameters are assumed to come from group-level distributions as indicated. The shaded bubbles represent observed variables (stimuli and responses), while unshaded bubbles represent quantities (parameters) to be estimated by the model. See Table 1 for further details.

relatively vague/uninformative priors (Table 1). We implemented the mcmc algorithm in R and in C++ using the Rcpp package (Eddelbuettel, François, Allaire, Ushey, Kou, Russel, Chambers, & Bates, 2021) to speed up the calculations. All this material and codes are provided in SM.

For each parameter of the model, we obtained 4 chains of 10,000 posterior samples, with a burn-in of 5000 samples and thinning factor of 5. Convergence of the Markov chains were diagnosed using a recent version of the R-hat proposed by Gelman and colleagues (Vehtari, Gelman, Simpson, Carpenter, & Bürkner, 2020) (criteria <1.01). For most group and subject level parameters we found convergence or slight departures from this criterion for some subjects and parameter values. Having checked convergence, for inferences we used the samples from one chain for each of the models run.

Model fit was assessed using posterior predictive checking (Gelman et al., 2014; Shiffrin, Lee, Kim, & Wagenmakers, 2008). Model comparison was made using Leave-One-Out Cross-Validation, and the Widely Applicable Information Criteria (WAIC), as both are recommended for hierarchical Bayesian models (Vehtari, Gelman, & Gabry, 2017), and are implemented in the loo package in R. The use of these measures requires comparing models with the same number of points. Because the two-cycle PEM is not defined at $p = 0.5$ (it produces a constant, with no error, see equations above), we excluded responses from this stimulus to make comparisons between the one- and two-cycle models (as in previous studies Rouder & Geary, 2014). In a second approach suggested by reviewers and explained in Results, we introduced a random middle that allowed us to address this issue in a more satisfactory way.

For the different steps of this work, we used the programs R (R Development Core Team, 2017) and MATLAB (The Mathworks, Inc., Natick, MA, USA). The codes and raw data will be provided in SM.

Table 1
Description of variables and parameters of PEM. The default values for the group-level parameters are the hyperpriors.

Variable	Default Value	Interpretation
$P_{i,j}$		Presented proportion for subject i at trial j
$Y_{i,j}$		Response proportion for subject i at trial j
$Z_{i,j}$	1 = internal, 2 = reversal, 3 = random guess	Indicator for types of responses for subject i at trial j
$\lambda_i = (\lambda_{i,int}, \lambda_{i,rev}, \lambda_{i,guess})$	$\lambda_{i,int} + \lambda_{i,rev} + \lambda_{i,guess} = 1$	Probabilities of each type of response for subject i in all the trails
β_i		Bias parameter for internal/reversal distributions for subject i
w_i		Noise parameter for internal/reversal distributions for subject i
$\alpha = (\alpha_1, \alpha_2, \alpha_3)$	$\alpha \sim Unif(0.01, 100)$	Group-level parameter of probabilities for each type of responses
μ_β	$\mu_\beta \sim Unif(0, 10)$	Group-level mean parameter for parameter β_i
σ_β^2	$\sigma_\beta^2 \sim Unif(0, 10)$	Group-level variance parameter for parameter β_i
μ_w	$\mu_w \sim Unif(0, 10)$	Group-level mean parameter for noise parameter w_i
σ_w^2	$\sigma_w^2 \sim Unif(0, 10)$	Group-level variance parameter for noise parameter w_i

3. Results

3.1. Proportion estimation task

For group level results, CPM fits, and a sample of individual subject variability please see Fig. 3 and text in the Introduction.

3.2. Proportion estimation model (PEM)

3.2.1. Model description

We built a mixture model of proportion estimation in line with the considerations above. We assumed that responses came from three sources: an internal representation of a proportion plus two contamination components: a reversal and guessing behavior. The internal component was more theoretically motivated and captured responses along the main diagonal in Fig. 3; the reversal and guessing behaviors were included more a posteriori, as part of data analysis.

To build the internal distribution, we assumed that subjects represent each magnitude composing the proportion stimuli (the sizes of red and blue columns in the proportion estimation task) as Gaussian activations that are mentally combined to form a proportion (Fig. 2). The resulting distribution of internal representation of proportions is shown in Fig. 5. The Gaussians composing the proportion were centered in the magnitudes of the stimuli, compressed or expanded by a certain amount (captured by the bias parameter β) and displayed scalar variability (their

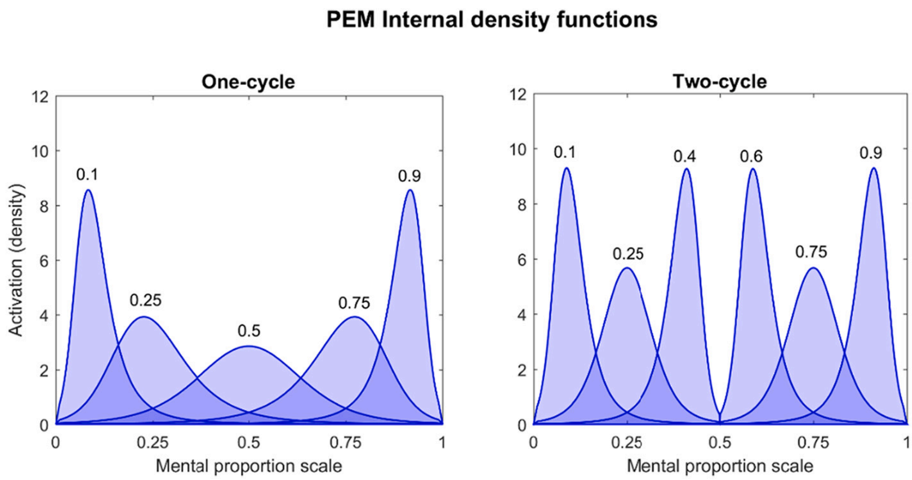

Fig. 5. Visualizing PEM. Internal density functions (int_p) for different presented proportions values (p) indicated on top of each hump. The parameters for the one- and two-cycle models are $\beta = 1$; $w = 0.4$. Note how the distributions for the one-cycle variant are progressively skewed as we move away from the middle ($p = 0.5$), and are symmetric around this point. These aspects were noted by Spence and Krizel (1994) and the model reproduces them, in line with our expectations. For the two-cycle, the distributions show the same behavior but they are compressed and repeated on each side of the half. In this model, the half is an anchor point and predicts no error.

standard deviations increased with their means). The rate of this increase is a measure of internal noise and was captured by the parameter w (see Methods for details).

Consistent with the formulation, we observe that the humps are pushed away from the half when $\beta > 1$, predicting underestimation for stimuli below 0.5 and overestimations above 0.5 (Fig. S1). Conversely, the humps move toward the middle when β values decrease below 1 (predicting overestimations of proportions below 0.5 and underestimation otherwise, consistent with the effects of β in CPM, see Fig. 1 and Fig. S1). On the other hand, the modulation of w causes a spread or shrinkage of the densities as the values of this parameter go up or down, respectively (see Fig. S1).

We then added to this central structure the two contaminating behaviors (reversal and guessing) after data inspection (see above and Fig. 6). Thus, each response for a presented proportion p and for each participant j, $Y_{j,p}$, was modeled as a sample from the mixture distribution:

$$Y_{j,p} \sim \lambda_{int\ j} \boldsymbol{INT}_p(\beta_j, w_j) + \lambda_{rev\ j} \boldsymbol{REV}_p(\beta_j, w_j) + \lambda_{guess\ j} \boldsymbol{Guess}_p$$

where INT_p, REV_p and $Guess_p$ are the internal and contaminating distributions, all with their respective weights. The reversal was derived from the internal function by just swapping the red and blue stimuli (i.e. p by $1\text{-}p$), and the guessing behavior was assumed to be uniformly distributed between 0 and 1 (see Figure6).

Paralleling the models of CPM, we developed one-cycle and two-cycle versions from these assumptions. and we framed the model as a hierarchical mixture model with subject and group level parameters (see Fig. 4 and details in Methods).

3.2.2. Fitting PEM

We implemented a Bayesian approach to fitting the model, as previous studies have found less variance in recovered parameters compared to using maximum likelihood in hierarchical models, especially in cases with small sample sizes (Farrel & Ludwig, 2008). Hierarchical models are flexible tools that combine information across different levels of analyses, providing a balance between entire pooling of the data (i.e. aggregated group analysis) or complete independence of subjects (no pooling; see Gelman et al., 2014). The Bayesian approach also helps incorporating information of previous studies in the form of priors. Note that in our case we used the Bayesian framework as a 'methodological tool' and not as part of the psychological components of the model (for a discussion of these distinctions see Lee, 2018; for Bayesian approaches to magnitude estimation see Petzschner, Glasauer, and Stephan (2015)).

Fig. 7 summarizes fit results. Joint posterior distributions as those shown help seeing potential interactions (correlations) between parameter estimates, which may be obscured when plotting only marginal posterior distributions (Shiffrin et al., 2008). In our case, we observed no relation between parameter estimates. At the group level, bias estimates from PEM one-cycle were slightly higher than 1 ($\mu_\beta = 1.074$ [1.029,1.12] (Mean, [95% credible interval]) and slightly smaller than 1 for the two-cycle model ($\mu_\beta = 0.949$ [0.874, 1.024]). These estimates were much closer to one (i.e. inferred less bias) compared to those obtained with CPM at group level shown above ($\beta_{mean} = 0.31$; $\beta_{median} = 1.4$), consistent with the model's capacity to absorb contaminating elements (see below). That is, using the traditional CPM approach, one will falsely find biased results where the CPM mistakes contamination for bias. In addition, the hierarchical framing of our model likely aids in absorbing part of the contaminating aspects (see below). Internal noise estimates were $\mu_w = 0.66$ [0.584, 0.746] for one-

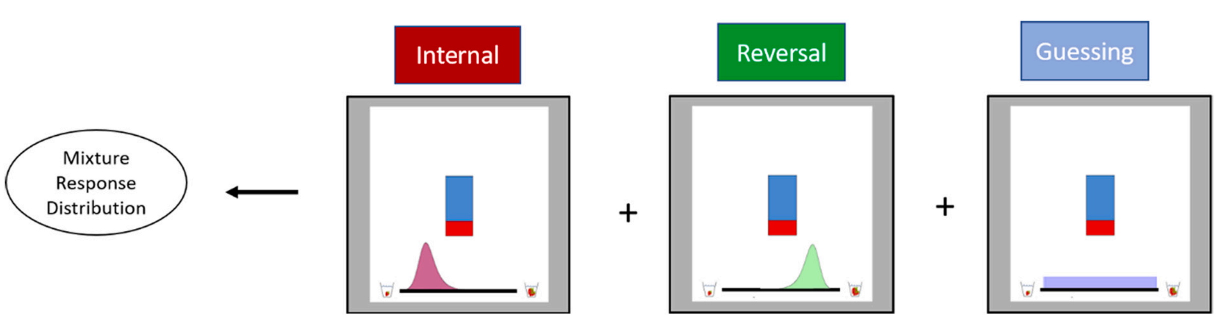

Fig. 6. Schematic representation of the mixture model. The humps along the response bars are the density functions of each of the components and for this stimulus ($p = 0.25$). These are hypothetical mental distributions (depicted for illustrative purposes) from which responses are sampled with probabilities proportional to the weights of each component. The parameters bias, internal noise, and weights have been omitted for clarity.

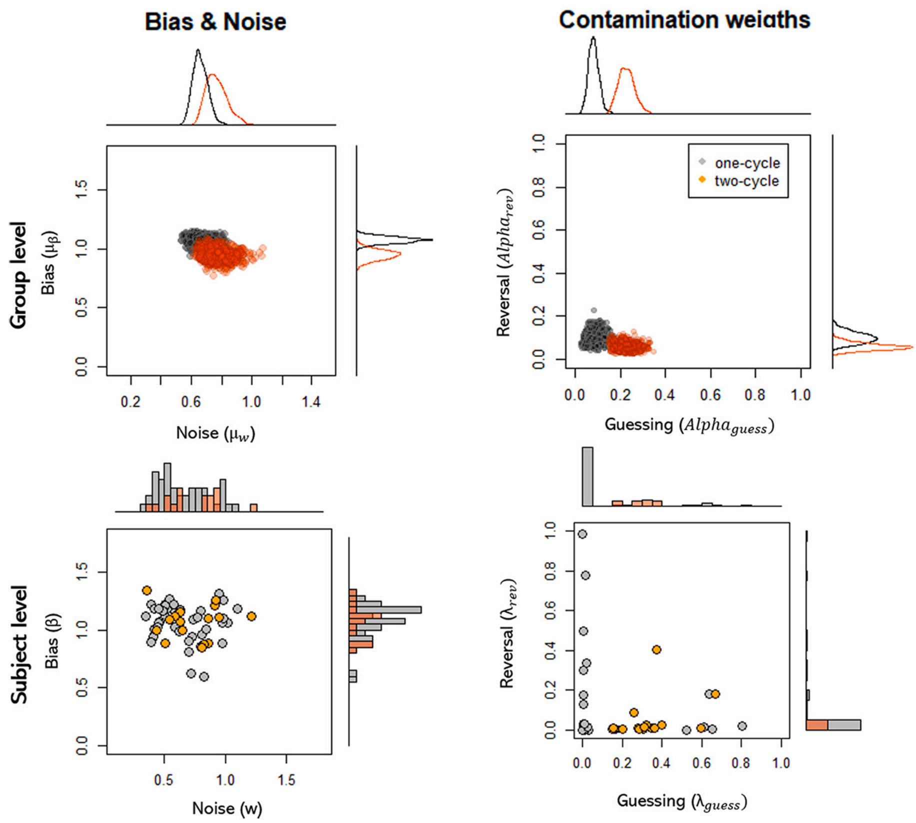

Fig. 7. Joint posterior distributions of group and subject level parameter. In group level panels (top row), dots are samples from the joint posterior distributions, with marginal densities on each side. In subject level panels, dots are the means of the posterior distributions for each subject, with their respective histograms.

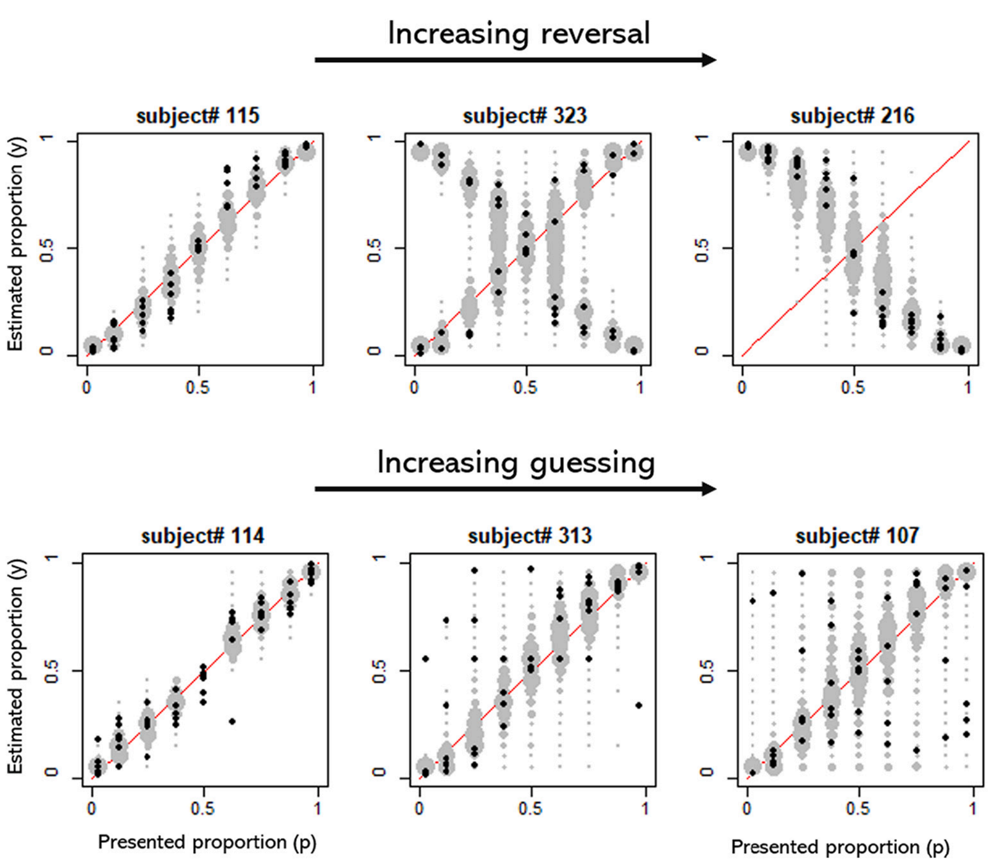

Fig. 8. Posterior predictive checking. The same sample of subjects shown in Fig. 3, with the predictions derived from their respective estimated parameters. Predictions (grey dots) were generated in two steps: 1) we sampled from the posteriors distributions of parameters, and 2) with these estimates, we obtained 'n' samples from the sampling distribution of the model (see equation in colors in the previous section). "n" was set equal to 54, the number of trials. We repeated this procedure 100 times and produced the bubble plots, superimposing the raw data (black dots). All subjects were best fit by a one-cycle PEM, except subject #114, who was best fit by a two-cycle PEM.

cycle and $\mu_w = 0.768$ [0.667, 0.905] for the two-cycle PEM. Contrary to our expectations, these values were considerably higher than those reported for discrimination of spatial length and area in children of the same age (Odic, 2017, reported values of *w* of 0.1 for both dimensions).

At the subject level, we found that most subjects were best fitted by the one-cycle PEM (46 out of 57 subjects [81%]; PEM two-cycle: 11 subjects [19%]). As shown in Fig. 7, we observed great variability in parameter estimates between subjects, especially in subjects best fitted by the two-cycle PEM. Bias estimates (posterior means) ranged between 0.6 and 1.317 for the one-cycle, and between 0.844 and 1.341 for the two-cycle. Internal noise estimates also overlapped between the models, ranging [0.342, 1.096] for the one-cycle, and [0.354, 1.207] for the two-cycle PEM, and still they were relatively high (as those of the group-level).

Regarding contaminating behaviors, we observed that reversal and guessing weights clustered around small values (see marginal histograms in Fig. 7) (PEM one-cycle: $alpha_{rev} = 0.101$ [0.063,0.147], $alpha_{guess} = 0.082$ [0.047,0.121], PEM two-cycle: $alpha_{rev} = 0.059$ [0.035,0.09], $alpha_{guess} = 0.227$ [0.172,0.289], but estimates covered the entire range between 0 and 1, suggesting extreme contaminating behaviors for some subjects.

To evaluate the performance of the model we implemented posterior predictive checking (Fig. 8). This procedure helps evaluating the descriptive adequacy of the model and the data (Shiffrin et al., 2008) and allows predicting future observations (Gelman et al., 2014). Overall, we observed good agreement between the model and the data, and that the different degrees of contamination we previously noted by visual inspection (Fig. 3) were clearly detected by the model. This suggests that the great variability between subjects resulted from 'genuine' variability in psychophysical estimates rather than on miscounting contaminating elements on the part of the model. In fact, when we fit the model ignoring contamination, we observed much dispersion on psychophysical estimates, especially for internal noise (Fig. 9).

Fit results suggested good performance of the model, which was able to account for different response patterns at the subject-level. However, further inspection of the data exposed a gap of the model. Specifically, we noticed that some subjects tend to be very precise when estimating proportions equal to the half ($p = 0.5$), even when their overall response pattern is captured by a one-cycle model (see for example subject#115 in Fig. 8). This is problematic for the model, because the one-cycle predicts the opposite behavior, i.e., increasing response variance toward the middle (see Figs. 5 and S2). The two-cycle version would seem to better account for this behavior. However, this model actually predicts no error at all at this point as it is conceived as a true anchor and so the function is undefined there (see how the data of subject#114 in Fig. 8 are not predicted in the middle; also see Fig. 5 and S2).

To tackle this issue, we introduced a simple modification to PEM, by assuming that responses from the middle can come from a different source than the rest of proportion estimations. We modeled this source as a Gaussian centered at the middle, and with standard deviation proportional to a parameter '*w_middle*'. The idea of appending a random middle was also suggested by Rouder & Geary (2014) to deal with the indefinability of their two-cycle model, but they did not assess it further (see Appendix in Rouder & Geary (2014)). Here we considered random middles for both the one-cycle and the two-cycle versions of PEM (Fig. S2 C,E). For the two-cycle, we included the corresponding possibility that subjects were more precise at the middle and the quarters (0.25, 0.75) (Fig. S2 D).

Thus, we fitted and compared 4 'variants' of PEM: two versions of one-cycle (the standard PEM "one-cycle" and with random middle ("one-cycle_H"), plus two versions of the two-cycle (with random middle ("Two-cycle_H), and with random middle and random quarters ("Two-cycle_H&Q")).[3] All have the same mixture structure but differed in the internal functions.

In line with the previous results, we found that most children were best fit by one-cycle models. Notably, we observed that more than half of those kids were best fit by the version with random middle ("one-cycle_H": 31 children [54%]; "one-cycle": 15 children [26%]). The remaining children were divided similarly between the two-cycle models ('two-cycle_H': 5 children [9%]; 'two-cycle_H&Q': 6 children [11%]). Also, consistent with the idea that estimates in the middle were more precise, we found that estimates of internal noise from the middle (*w_middle*) were significantly smaller than those obtained from the rest of the proportion values (model: **"one-cycle_H"**: $\mu_{w\_middle} = 0.312$ [0.242,0.388]; $\mu_w = 0.671$ [0.604, 0.747]; model: **"two-cycle_H"**: $\mu_{w\_middle} = 0.294$ [0.255,0.372]; $\mu_w = 1.24$ [1.043, 1.489]); **"two-cycle_H&Q"**: $\mu_{w\_middle} = 0.496$ [0.447,0.55]; $\mu_w = 1.006$ [0.874, 1.143]).

These results suggest that adding an independent source of noise to the middle (and the quarters) provide a better description of the data than using the standard PEM for some subjects. This was also confirmed by posterior predictive checking (see Fig. 10).

3.3. Applying PEM: Assessing scaling effects on proportion estimation

In this last section we highlighted the value of using our model to the study of scaling effects in proportional reasoning. In a proportional reasoning task, the scaling factor tracks the degree of disparity between the two spaces put in correspondence (the multiplicative relationship – similar to our example of a map being related to the world by a scaling factor). Recent studies have shown that subjects are sensitive to this factor, displaying higher errors as the scaling increases (Boyer & Levine, 2012; Möhring et al., 2015; Vasilyeva & Huttenlocher, 2004). The typical approach would look at the overall error – but does an increase in error suggest a change in bias, in internal noise, in guessing, or in all three? Do overall patterns of responses change across scaling levels so that different PEM variants better describe them?

To study this, we varied the ratio between the total height of the columns and the width of the response line (as in Möhring et al., 2015) in three levels (Fig. 11A). As the width of the response line was held fixed, stimuli got progressively smaller as scaling factor increased, thus requiring scaling up the stimuli.

As shown in Fig. 11A, we found that the percentage of subjects best fit by the different versions of PEM varied across scaling factor levels. A noteworthy aspect was the increase in the percentage of children best fit by the one-cycle model as scaling increased. Interestingly, we observed that many subjects were best fit by the same model across scaling1:2 and scaling 1:3, while others displayed more variability in their selected model.

Regarding parameter estimates, we focused on the fits of the one-cycle and two-cycle_H&Q for illustrative purposes. For both models, we observed a trend of increase in the population bias and noise parameters with scaling levels (Fig. 11C). The amount of guessing and reversal did not vary across scaling levels (see how the curves overlap in Fig. 11C). The modulations of μ_w are in line with previous reports showing that estimation errors increase with scaling factor levels (Vasilyeva & Huttenlocher, 2004). Interestingly, the specific values μ_β, which increased from around $\mu_\beta = 1$ for scaling 1:1 to values above 1, indicate that responses were progressively biased and "pushed away" from the reference points (the middle and the quarters) as stimuli became smaller and smaller. These values ($\beta s > 1$) are consistent with an expansion of the signal as stimuli were scaled up. This subtle behavior

[3] We did not include the two-cycle PEM without random middle in this comparison precisely because it has the problem of not fitting estimations from the middle and so it includes less points.

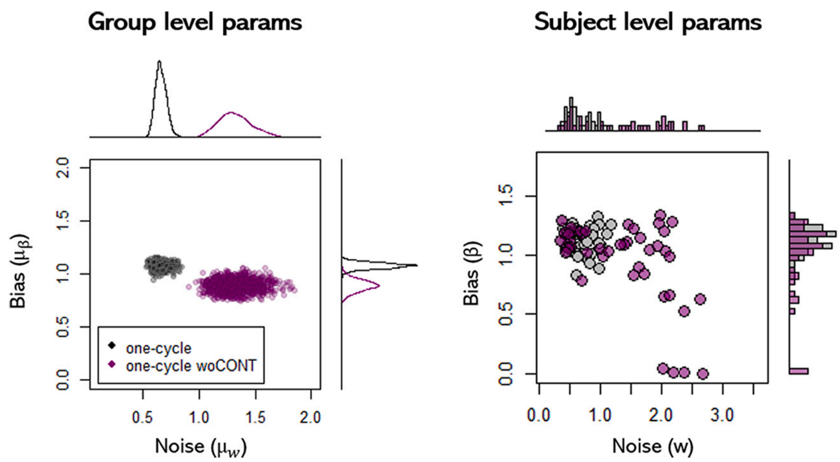

Fig. 9. Effects of ignoring contaminating elements on parameter estimates. Shown are subjects best fit by one-cycle PEM with and without contaminating behaviors. Note the larger scale in subject-level noise estimates.

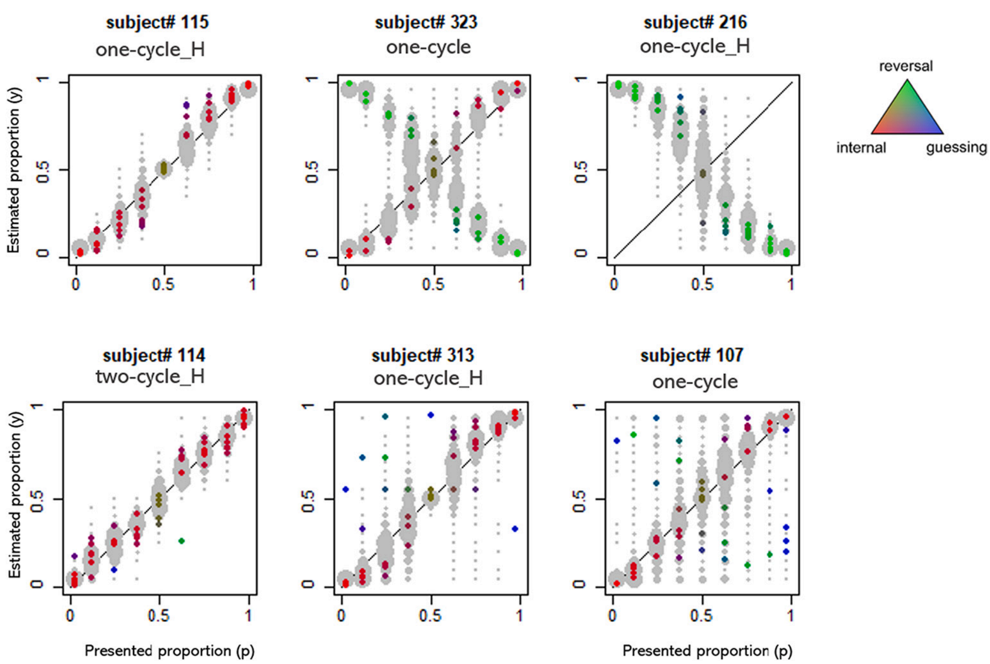

Fig. 10. The same sample of subjects as in Fig. 8. Predictions are now from the different variants of PEM, as indicated on top of each plot. Compare the fits from subjects 115, 114 and 313 with those from Fig. 8; now they show much better adequacy. Also, individual trials have been colored ('filtered') according to their likelihood values of PEM distributions: internal, reversal and guessing. The parameter estimates (β and w) used for the internal and reversal densities are the posterior means obtained from the best fit variant of PEM for each subject.

would not have been evident had we focused on overall measures of error.

4. Discussion

In this study we developed a model of proportion estimation, that extends previous approaches in principled and methodological ways. Following the generative logic of CPM, we derived a statistical description of the internal representation of a proportion, with bias and noise parameters directly linking absolute with relative magnitudes. Methodologically, we introduced a mixture of components to quantify contaminating behaviors (reversal and guessing) and framed the model in a hierarchical way. We found that the model could mimic important aspects of the variability of responses (skewedness) and displayed good descriptive adequacy with the data, capturing commonalities but also great variations between subjects in terms of psychophysical estimates and contaminating behaviors. Finally, we reported consistent modulations of psychophysical parameters with spatial scaling levels, providing insight on an important phenomenon in proportional reasoning.

In building the model, we made two central assumptions: that subjects mentally computed proportions of noisy representations, and that such individual representations were described by Gaussians with scalar variability. To what extent did we find support for these assumptions, and for the model more generally? On the one hand, we found empirical support for the model in that the shape of the internal density has the same asymmetries (skewedness) that had been previously observed in estimation data (e.g. Spence, 1990), thus mimicking an important empirical aspect of response variability; Spence and Krizel (1994) explained the skewedness by proposing repulsion forces that push responses away from the boundaries. The existence of such forces was attributed to the use of imperfect strategies on the part of subjects but as noted by Hollands and Dyre (2000), the nature of such forces was not independently verified. More importantly, such a theory is incompatible with the observation that the bulk of responses can move *toward* the boundaries (as when $\beta > 1$) while also displaying skewedness. Our model was able to reproduce these patterns by just assuming that subjects mentally combined noisy internal representations of absolute magnitudes as mental proportions. Additionally, we observed good

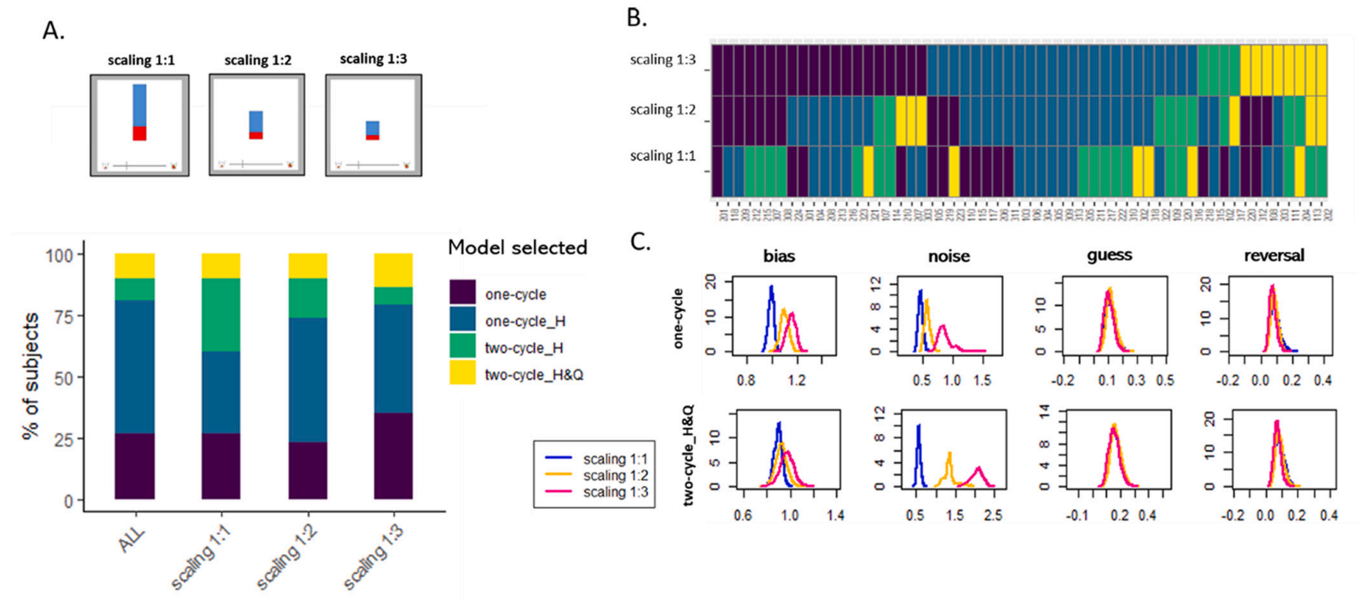

Fig. 11. Effects of scaling factor on model selection and parameter estimates. A, proportion of subjects that were best fit by the different models. On top of each column, a stimulus depicting scaling factor level is displayed. B, models selected for each participant and scaling factor level. C, Posterior distribution of group-level parameters for the one-cycle and two-cycle_H&Q models. Note how the curves are progressively shifted to the right for bias and noise parameters, while they overlap for contaminating weights.

performance of the model in describing various aspects of the data, which together provide empirical support for the model. And, we can answer these questions by checking whether the predictions of the model (derived from the assumptions) were met or not.

Due to the simple generative framework, we expected to find similar levels of bias and internal noise as those reported for absolute magnitudes. Indeed, although our results showed great variability in bias estimates between subjects, group-level estimates yield values comparable to those reported in previous studies for the perception of absolute areas and lengths (values respectively near 0.8 and close to 1; Teghtsoonian, 1965), in support of the model. These values were similar to those reported by Spence and Krizel (1994) (β = 0.8 at group level analysis), who implemented a very similar task to ours (in kids of a similar age, but framed as a graph-reading task and not as a juice task as in our case). However, internal noise estimates concentrated around much higher values ($w = 0.6$) than those reported for absolute areas and lengths ($w = 0.1$, as observed by Odic (2017) in children of the same age). These discrepancies cast doubts on the model's tenet of linking absolute with relative magnitudes through internal noise, yet we believe that methodological differences can account for them.

Indeed, while Odic (2017) used irregular blobs as stimuli for area judgments here we used rectangular shapes. This difference may matter, considering the results of Morgan (2005) who found higher weber fractions than Odic for the discrimination of areas using rectangles (although smaller than ours) but in adult participants. Given that these were adults participants, we would expect them to have smaller w's than our kids, just like for the development of approximate number precision (Halberda & Feigenson, 2008; Halberda, Ly, Wilmer, Naiman, & Germine, 2012). So, in this scenario, our estimates might be accurate. Alternatively, there may be true differences in noise estimates arising from the tasks implemented (estimation versus discrimination tasks) (Guillaume, Gevers, & Content, 2015), regardless of whether we compare absolute or relative magnitudes. In this scenario, the assumption of the model linking absolute with relative magnitudes would be valid but possibly restricted to estimation tasks. Follow up studies using proportion estimation of blob figures and regular shapes, combined with absolute estimation and discrimination tasks could shed additional light on this issue.

The above considerations suggest that our model is sufficient to explain proportion estimation data, but are these model assumptions *necessary*? In other words, could we think of a different mental operation or a different source of noise and obtain an equally satisfactory model? This is a harder issue to evaluate, because it implies acknowledging what constitutes a 'good model' besides its capacity to fit data, and also because the space of candidate models can be large (Lewandowsky & Farrel, 2011). We assumed a Gaussian model with scalar variability to structure the representation of individual magnitudes, because it provided a relatively simple yet well supported account to build the model, but other models are possible. We discarded the log-Gaussian model (Dehaene, 2007), which assumes equal-variance normal activations on a log internal scale, because it can only account for compressive biases that push proportion estimations toward the middle, but it cannot account for the reverse situation (i.e. expansive biases that push estimations away from the middle). Furthermore, recent evidence suggests that the power function, but not logarithmic functions, can properly describe sensory and neural processes that transform physical scales into internal scales (Pardo-vazquez et al., 2019). On the other hand, Laming (1997)'s χ^2 (chi-square) model of magnitude representation, which has received relatively less consideration in the literature could be an interesting alternative to explore in future studies. Especially, because the Beta distribution (widely used in statistics to model random variables bounded between 0 and 1) can be derived from the proportion of χ^2 distributions, thus potentially offering a compelling model for the internal representation to contrast with our account. However, the seemingly unclear substantive interpretation of shape parameters in the χ^2 model, still warrant further study to make this an interesting idea from a psychological point of view.

One potential limitation of our study is that our model did not explicitly consider potential contributions of memory, attention and motor processes that may also contribute to generate variability and biases in estimation behavior (Lockhead, 2004; Odic et al., 2016; Petzschner et al., 2015; Shepard, 1981; R. Teghtsoonian, 2012; Wei & Stocker, 2015; Zax, Williams, et al., 2019). While these factors can be incorporated in further elaborations of the model, our formulation is not incompatible with an influence of these factors as manifested in the psychophysical parameter. Thus, while it is implicit in our representational model that internal noise and bias result mainly from sensory measurement processes (Odic et al., 2016; Pardo-vazquez et al., 2019), we avoided adopting a strong interpretation of the psychological meaning of w and β parameters, which likely represent a composite of processes rather than a monolithic aspect of mental representations (see the above cited papers), and accordingly presented the model in a general way.

Furthermore, the fact that we implemented a line as the response scale renders the assumption that it contributed zero bias plausible, since estimation studies have shown that the perception of lengths is close to veridical (M. Teghtsoonian, 1965), thus suggesting that the bias (and noise) may arise from the perception of stimuli and/or from the mapping onto the response scale (see Shepard, (1981) for a thorough treatment and interpretation of bias in cross-modal matching tasks).

A second, related limitation of our study, is that we implemented only one proportion estimation task to empirically evaluate the model which may limit the generality of our conclusions. This may be valid with respect to the parameter estimates we obtained, and to our ability to generalize our results to other studies/tasks, but it does not obfuscate the theoretical validity of the model itself. Moreover, as CPM can be seen as a special case of our account, the empirical validity of CPM is automatically inherited by our model. In contrast, the task we chose has some noteworthy advantages (some discussed by Spence & Krizel, 1994), for example, it is fairly intuitive for children as it does not require an explicit knowledge of symbolic proportions (i.e. or fractions like 3/5) and the specific stimuli we used (continuous bars) reduce the likelihood of inducing a whole-number bias. This bias broadly refers to the wrong use of whole number strategies in fraction-related problems (Siegler et al., 2013), and in our context applies to the potential focus on only one piece of information (e.g. numerator) during the proportion estimation task. Studies have shown that the use of discrete stimuli (e.g. a collection of balls) are more likely to induce this bias than using continuous stimuli as the ones employed here (Boyer et al., 2008; Jeong, Levine, & Huttenlocher, 2007). Finally, recent studies have successfully used this task to uncover links between the accuracy in estimating proportions with formal math abilities in children (Gouet et al., 2020; Möhring et al., 2015), suggesting wider applications of this task to connect with other fields in cognitive research.

An interesting result showed up when we introduced the new variants of PEM to account for the high precision of estimations in the middle. This aspect was highlighted by reviewers and we attempted a simple solution, following similar previous attempts (see Rouder & Geary, 2014). Remarkably, we found that a great proportion of subjects were best fit by the variant of one-cycle that predicts precise estimations in the middle. This means that some subjects were very precise in the middle but still do not 'use' it as a virtual anchor (as is the case for those who were best fit by two-cycle model). Developmental studies have shown that the 'half' (Spinillo & Bryant, 1991) and the ability of mental split (Confrey, 1994) play crucial roles in early proportional reasoning abilities (Spinillo & Bryant, 1991), and that children progressively use the middle as an anchor point (e.g. Möhring et al., 2018; Slusser et al., 2013). Speculating, we think that the group of subjects just referred to may be in an intermediate state between one-cycle and two-cycle behavior (see also (Zax, Slusser, & Barth, 2019). Future studies with a wider age range could further elaborate on the substantive validity of the variants of PEM, and refinements could be incorporated, for example in the form of priors that give more weight to the representations of the middle, under the efficient coding and Bayesian scheme of Wei and Stocker (2015) or through category models (Huttenlocher et al., 1991).

In the last part of our work, we sought to illustrate the utility of our model to the study of scaling effects. The nature of the scaling process is not yet clear, but it is thought to involve some analog transformations that mentally shrink or expand the representations generated from one space to be equivalent to those from another space (Boyer & Levine, 2012; Möhring et al., 2018; Vasilyeva & Huttenlocher, 2004). Our results give support to this idea, in showing a monotonic increase of bias and internal noise with scaling levels, consistent with an expansion of the internal signal as stimuli had to be scaled-up. These results find nice parallels with results from the literature of probability and frequency judgment. As referred to in the Introduction, different studies have found s-shaped pattern of biases when subjects estimate the frequency of discrete events or elements (e.g. Varey et al., (1990), reviewed in H. Zhang & Maloney, 2012). Investigating the factors that modulate biases in probability/frequency estimation, H. Zhang and Maloney (2012) noted from their results and others that as the sample size from which estimations are made gets smaller, small frequencies are underestimated and high frequencies are over estimated (similar to what we observe here when the columns get progressively smaller with scaling levels). It will be interesting to test whether concomitant modulation of internal noise can also be inferred in these frequency/probability judgments and decision-making, also incorporating dynamical aspects of probability perception (Gallistel, Krishan, Liu, Miller, & Latham, 2014). Paraphrasing Zhang and Maloney (2012), having developed the tools for measuring bias and internal noise, the key question is to understand what are the functional implications of these psychophysical signatures.

We admit that, in retrospect, it seems obvious that one cannot arrive at an accurate understanding of the psychology that supports the successful performance of a task without having a model that successfully fits the data in its most raw form. It is in the raw form that all of the nuance of performance is laid bare. And, perhaps, proportion estimation is an interesting task precisely because of the subtleties in response profiles and the links from internal representations of absolute magnitudes (i.e., lengths) to proportions (i.e., ratios of lengths). PEM is one attempt to take these considerations seriously.

Declaration of Competing Interest

None.

Acknowledgements

We thank all the parents and children who participated in the study. We thank to Sylvia Campos for contacting us to Villa Macul School and to the school's staff. We acknowledge the Comisión Nacional de Ciencia y Tecnología (CONICYT) for the PhD scholarship (# 21130131) to C.G, the project Fondecyt #1141040 to M.P. and the James S. McDonnell Scholar Award to J.H. Thanks to Tyler Knowlton and Josh Langfus for data analysis assistance, and thanks to Raul Gouet for an insight on calculations. We thank the anonymous reviewers for their careful reading and suggestions to improve earlier versions of the manuscript.

Finally, we want to express our gratitude with Jacques Mehler for his inspiring ideas, and for building a community around the warm pyre of cognition.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2021.104805.

References

Acredolo, C., Connor, J. O., Banks, L., & Horobin, K. (1989). *Children's ability to make probability estimates: Skills revealed through application of Anderson's functional measurement methodology* (pp. 933–945).

Barth, H. C., & Paladino, A. M. (2011). The development of numerical estimation: Evidence against a representational shift. *Developmental Science, 14*(1), 125–135.

Boyer, T. W., & Levine, S. C. (2012). Child proportional scaling: Is 1/3=2/6=3/9=4/12? *Journal of Experimental Child Psychology, 111*(3), 516–533. https://doi.org/10.1016/j.jecp.2011.11.001.

Boyer, T. W., & Levine, S. C. (2015). Prompting children to reason proportionally : Processing discrete units as continuous amounts. *Developmental Psychology, 51*(5), 615–620.

Boyer, T. W., Levine, S. C., & Huttenlocher, J. (2008). Development of proportional reasoning : Where young children go wrong. *Developmental Psychology, 44*(5), 1478–1490.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York: Springer.

Chen, Q., & Verguts, T. (2017). Numerical proportion representation : A Neurocomputational account. *Frontiers in Human Neuroscience, 11*, 1–8. https://doi.org/10.3389/fnhum.2017.00412.

Confrey, J. (1994). Splitting, similarity, and the rate of change: A new approach to multiplication and exponential functions. Albany: State Uni. In G. Harel, & J. Confrey (Eds.), *The development of multiplicative reasoning in the learning of mathematics* (pp. 291–330).

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

Development Core Team, R. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org/.

Eddelbuettel D, François R, Allaire J, Ushey K, Kou Q, Russel N, Chambers J, Bates D (2021). Rcpp: Seamless R and C++ Integration. R package version 1.0.6, URL http://CRAN.R-Project.org/package=Rcpp.

Farrel, S., & Ludwig, C. J. H. (2008). Bayesian and maximum likelihood estimation of hierarchical response time models. *Psychonomic Bulletin & Review, 15*(6), 1209–1217. https://doi.org/10.3758/PBR.15.6.1209.

Gallistel, C. R., Krishan, M., Liu, Y., Miller, R., & Latham, P. E. (2014). The Perception of Probability. *Psychological Review, 121*(1), 96–123. https://doi.org/10.1037/a0035232.

Garrett, D. D., Kovacevic, N., Mcintosh, A. R., & Grady, C. L. (2011). The importance of being variable. *The Journal of Neuroscience, 31*(12), 4496–4503. https://doi.org/10.1523/JNEUROSCI.5641-10.2011.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian Data Analysis* (3rd ed.). Taylor & Francis.

Gouet, C., Carvajal, S., Halberda, J., & Peña, M. (2020). Training non-symbolic proportional reasoning transfers to symbolic proportional reasoning in 4th grade children. *Cognition, 197*. https://doi.org/10.1016/j.cognition.2019.104154.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Grondin, S., & Killeen, P. (2009). Tracking time with song and count : Different weber functions for musicians and nonmusicians. *Attention, Perception, & Psychophysics, 71* (7), 1649–1654. https://doi.org/10.3758/APP.

Guillaume, M., Gevers, W., & Content, A. (2015). 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. (2016). Epistemic Limitations and Precise Estimates in Analog Magnitude Representation. In D. Barner, & A. S. Baron (Eds.), *Core Knowledge and Conceptual Change* (pp. 1–24). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190467630.003.0010.

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, 109*(28), 11116–11120. https://doi.org/10.1073/pnas.1200196109.

Halberda, J., & Odic, D. (2015). The precision and internal confidence of our approximate number thoughts. In D. C. Geary, D. B. Berch, & K. M. Koepke (Eds.), *Evolutionary origins and early development of basic number processing* (pp. 305–333). Academic Press. https://doi.org/10.1016/B978-0-12-420133-0.00012-0.

Hoff, P. D. (2009). *A first course in Bayesian statistical*. Springer.

Hollands, J. G., & Dyre, B. P. (2000). Bias in proportion judgments: The cyclical power model. *Psychological Review, 107*(3), 500–524. https://doi.org/10.1037//0033-295X.107.3.500.

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review, 98*(3), 352–376.

Inglis, M., & Gilmore, C. (2014). Acta Psychologica indexing the approximate number system ☆. *Acta Psychologica, 145*, 147–155. https://doi.org/10.1016/j.actpsy.2013.11.009.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. Basic Books, Inc.

Jacob, S. N., Vallentin, D., & Nieder, A. (2012). Relating magnitudes: The brain's code for proportions. *Trends in Cognitive Sciences, 16*(3), 157–166. https://doi.org/10.1016/j.tics.2012.02.002.

Jeong, Y., Levine, S. C., & Huttenlocher, J. (2007). The Development of proportional reasoning : Effect of continuous versus discrete quantities. *Journal of Cognition and Development, 8*(2), 237–256.

Kemp, C., Bernstein, A., & Tenenbaum, J. B. (2005). A generative theory of similarity. In G. B. Bara, L. W. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th annual conference of the cognitive science society*. NJ:Erlbaum: Mahwah.

Laming, D. (1997). *The measurement of sensation*. Oxford, UK: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198523420.001.0001.

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology, 55*(1), 1–7. https://doi.org/10.1016/j.jmp.2010.08.013.

Lee, M. D. (2018). Bayesian methods in cognitive modeling. In J. T. Wixted (Ed.), *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience (Fourth)*. John Wiley and Sons. https://doi.org/10.1002/9781119170174.epcn502.

Lee, M. D., Gluck, K. A., & Walsh, M. M. (2016). *Understanding the complexity of simple decisions: Modeling multiple behaviors and switching strategies*.

Lee, M. D., & Sarnecka, B. W. (2009). A model of knower-level behavior in number concept development. *Cognitive Science, 34*, 51–67. https://doi.org/10.1111/j.1551-6709.2009.01063.x.

Lesh, R., Post, T., & Behr, M. (1988). Proportional reasoning. In J. Hiebert, & M. Behr (Eds.), *Number concepts and operations in the middle grades* (pp. 93–118). Reston: Lawrence Erlbaum & National Council of Teachers of Mathematics.

Lewandowsky, S., & Farrel, S. (2011). *Computational modeling in cognition: Principles and practice*. SAGE Publications, Inc.

Lewis, M. R., Matthews, P. M., & Hubbard, E. M. (2015). Neurocognitive architectures and the nonsymbolic foundations of fractions understanding. In D. B. Berch, D. C. Geary, & K. M. Koepke (Eds.), *Development of mathematical cognition: Neural substrates and genetic influences* (pp. 141–160). San Diego, CA: Academic Press.

Lockhead, G. R. (2004). Absolute judgments are relative: A reinterpretation of some psychophysical ideas. *Review of General Psychology, 8*(4), 265–272. https://doi.org/10.1037/1089-2680.8.4.265.

Lourenco, S. F., & Bonny, J. W. (2016). Representations of numerical and non-numerical magnitude both contribute to mathematical competence in children. *Developmental Science*, 1–16. https://doi.org/10.1111/desc.12418.

Matthews, P. G., Lewis, M. R., & Hubbard, E. M. (2016). Individual differences in nonsymbolic ratio processing predict symbolic math performance. *Psychological Science, 27*(2), 191–202. https://doi.org/10.1177/0956797615617799.

Möhring, W., Frick, A., & Newcombe, N. (2018). Spatial scaling, proportional thinking, and numerical understanding in 5- to 7-year old children. *Cognitive Development, 45*, 57–67. https://doi.org/10.1016/j.cogdev.2017.12.001.

Möhring, W., Newcombe, N. S., Levine, S. C., & Frick, A. (2015). Spatial proportional reasoning is associated with formal knowledge about fractions. *Journal of Cognitive Development.*. https://doi.org/10.1080/15248372.2014.996289.

Morgan, M. J. (2005). The visual computation of 2-D area by human observers. *Vision Research, 45*(19), 2564–2570. https://doi.org/10.1016/j.visres.2005.04.004.

Nakajima, Y. (1987). A model of empty duration perception. *Perception, 16*, 485–520.

Nieder, A., & Dehaene, S. (2009). Representation of number in the brain. *Annual Review of Neuroscience, 32*, 185–208. https://doi.org/10.1146/annurev.neuro.051508.135550.

Odic, D. (2017). Children's intuitive sense of number develops independently of their perception of area, density, length, and time. *Developmental Science, (September 2016)*, 1–15. https://doi.org/10.1111/desc.12533.

Odic, D., Im, H. Y., Eisinger, R., Ly, R., & Halberda, J. (2016). PsiMLE: A maximum-likelihood estimation approach to estimating psychophysical scaling and variability more reliably, efficiently, and flexibly. *Behavior Research Methods, 48*(2), 445–462. https://doi.org/10.3758/s13428-015-0600-5.

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.

Pardo-vazquez, J. L., Saa, J. R. C., Valente, M., Damião, I., Costa, T., Vicente, M. I., … Renart, A. (2019). The mechanistic foundation of weber ' s law. *Nature Neuroscience, 22*(9), 1493–1502. https://doi.org/10.1038/s41593-019-0439-7.

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

Piazza, M., 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 (New York, N.Y.), 306*(5695), 499–503. https://doi.org/10.1126/science.1102085.

Plantinga, J., & Trainor, L. J. (2005). Memory for melody: Infants use a relative pitch code. *Cognition, 98*(1), 1–11. https://doi.org/10.1016/j.cognition.2004.09.008.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychological Bulletin & Review, 9*(3), 438–481.

Rice, J. A. (2007). *Mathematical statistics and data analysis* (3th ed.). Thomson Brooks/Cole.

Rouder, J. N., & Geary, D. C. (2014). Children' s cognitive representation of the mathematical number line. *Developmental Science, 4*, 525–536. https://doi.org/10.1111/desc.12166.

Shepard, R. N. (1981). Psychological relations and psychophysical scales: On the status of "direct" psychophysical measurement. *Journal of Mathematical Psychology, 24*(1), 21–57. https://doi.org/10.1016/0022-2496(81)90034-1.

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science, 32*, 1248–1284. https://doi.org/10.1080/03640210802414826.

Siegler, R., Fazio, L. K., Bailey, D. H., & Zhou, X. (2013). Fractions: The new frontier for theories of numerical development. *Trends in Cognitive Sciences, 1*, 13–19.

Siegler, R., & Vago, S. (1978). *The development of a proportionality concept: judging relative fullness. 395* pp. 371–395).

Slusser, E. B., Santiago, R. T., & Barth, H. C. (2013). Developmental change in numerical estimation. *Journal of Experimental Psychology: General, 142*(1), 193–208.

Sophian, C. (2000). Perceptions of proportionality in young children : Matching spatial ratios. *Cognition, 75*, 145–170.

Spence, I. (1990). Visual psychophysics of simple graphical elements. *Journal of Experimental Psychology: Human Perception and Performance, 16*(4), 683–692. https://doi.org/10.1037/0096-1523.16.4.683.

Spence, I., & Krizel, P. (1994). Children' s perception of proportion in graphs. *Child Development, 65*(4), 1193–1213. https://doi.org/10.1111/j.1467-8624.1994.tb00812.x.

Spinillo, A. G., & Bryant, P. (1991). Children's proportional judgments: The Importanee of "Half". *Child Development, 62*, 427–440.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review, 64*(3), 153–181. https://doi.org/10.1037/h0046162.

Teghtsoonian, M. (1965). The judgment of size. *The American Journal of Psychology, 78* (3), 392–402.

Teghtsoonian, R. (2012). The standard model for perceived magnitude : A framework for (almost) everything known about it. *Everything Known About It, 125*(2), 165–174.

Varey, C. A., Mellers, B. A., & Birnbaum, M. H. (1990). Judgments of proportions. *Journal of Experimental Psychology. Human Perception and Performance, 16*(3), 613–625. https://doi.org/10.1037/0096-1523.16.3.613.

Vasilyeva, M., & Huttenlocher, J. (2004). Early Development of Scaling Ability. *40*(5), 682–690. https://doi.org/10.1037/0012-1649.40.5.682.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2020). *Rank-normalization, folding, and localization: An improved R-hat for assessing convergence of MCMC* (pp. 1–27).

Vukovic, R. K., Fuchs, L. S., Geary, D. C., Jordan, N. C., Gersten, R., & Siegler, R. S. (2014). Sources of individual differences in children's understanding of fractions. *Child Development, 00*(0), 1–16.

Wei, X., & Stocker, A. A. (2015). A Bayesian observer model constrained by efficient coding can explain "anti-Bayesian" percepts. *Nature Neuroscience, October*, 1–11. https://doi.org/10.1038/nn.4105.

Whalen, J., Gallistel, C. R., & Gelman, R. (1999). Nonverbal counting in humans: The psychophysics of number representation. *Psychological Science, 10*, 130–137. https://doi.org/10.1111/1467-9280.00120.

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in python. *Frontiers in Neuroinformatics, 7*(August), 1–10. https://doi.org/10.3389/fninf.2013.00014.

Zax, A., Slusser, E., & Barth, H. (2019). Journal of experimental child spontaneous partitioning and proportion estimation in children ' s numerical judgments. *Journal of Experimental Child Psychology, 185*, 71–94. https://doi.org/10.1016/j.jecp.2019.04.004.

Zax, A., Williams, K., Patalano, A. L., Slusser, E., Barth, H., Zax, A., … Slusser, E. (2019). What do biased estimates tell us about cognitive processing? Spatial judgments as proportion estimation. *Journal of Cognition and Development, 20*(5), 702–728. https://doi.org/10.1080/15248372.2019.1653297.

Zeigenfuse, M. D., & Lee, M. D. (2010). A general latent assignment approach for modeling psychological contaminants. *Journal of Mathematical Psychology, 54*, 352–362.

Zhang, H., & Maloney, L. T. (2012). Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in Neuroscience, 6*(January), 1–14. https://doi.org/10.3389/fnins.2012.00001.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453*, 233–236. https://doi.org/10.1038/nature06860.