Cognition 252 (2024) 105918

Contents lists available at ScienceDirect

Cognition

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

Continuous and discrete proportion elicit different cognitive strategies

Michelle A. Hurst [a,*], Steven T. Piantadosi [b]

[a] Rutgers University, New Brunswick
[b] University of California, Berkeley

ARTICLE INFO

Keywords:
Proportion
Strategy
Bayesian analysis
Model comparison

ABSTRACT

Despite proportional information being ubiquitous, there is not a standard account of proportional reasoning. Part of the difficulty is that there are several apparent contradictions: in some contexts, proportion is easy and privileged, while in others it is difficult and ignored. One possibility is that although we see similarities across tasks requiring proportional reasoning, people approach them with different strategies. We test this hypothesis by implementing strategies computationally and quantitatively comparing them with Bayesian tools, using data from continuous (e.g., pie chart) and discrete (e.g., dots) stimuli and preschoolers, 2nd and 5th graders, and adults. Overall, people's comparisons of highly regular and continuous proportion are better fit by proportion strategy models, but comparisons of discrete proportion are better fit by a numerator comparison model. These systematic differences in strategies suggest that there is not a single, simple explanation for behavior in terms of success or failure, but rather a variety of possible strategies that may be chosen in different contexts.

1. Introduction

Proportional information is ubiquitous in our everyday lives and has a prominent role in many cognitive theories, where people's ability to make probabilistic inferences from proportional information is a central learning mechanism (Denison and Xu, 2012; Xu, 2019). Yet, there is not an accepted, standard account of how people reason about proportion. Part of the difficulty is that there are many surprising and even contradictory claims. For example, thinking about proportional information has been called both inherently difficult (Lamon, 1993) and intrinsically simple (Gillard et al., 2009). It also has a surprising developmental trajectory, with researchers finding intuitive proportional reasoning in infancy (e.g., Denison et al., 2013; Denison and Xu, 2010; Xu and Denison, 2009; Xu and Garcia, 2008; although see Placì et al., 2020; Téglás et al., 2011) but profound difficulties in older children and adults (e.g., Boyer et al., 2008; Bryant and Nunes, 2012; Fazio et al., 2016; Girotto et al., 2016; Hurst et al., 2021; Piaget and Inhelder, 1975; Schneider and Siegler, 2010; Tversky and Kahneman, 1974).

We argue that these patterns and inconsistencies have a simple origin: variation in behavior across tasks and development reflects the use of different strategies, some of which may be entirely non-proportional. Though *we* see all these tasks as involving the same underlying concept – proportion – participants interpret the stimuli in fundamentally different ways across tasks and ages, resulting in the use of distinct context-dependent strategies. In the current paper, we test this proposal using a Bayesian model framework which allows us to formalize and quantitatively compare competing strategies as explanations of behavior. Specifically, we use model comparison to investigate different patterns in strategy use (i.e., relative model fit) across tasks with different stimuli formats.

This idea of people using different strategies across contexts is not a new one. Siegler has shown in many different domains and across development that people have access to many different strategies for the same task (e.g., Siegler, 1987; Siegler, 1991, 1994). For example, in the domain of symbolic fractions, a mathematics topic deeply related to proportion, self-report data suggest that people use a mix of strategies, including some that rely on incomplete or incorrect information (Fazio et al., 2016). Our goal here is to formalize the proposal that variation in behavior originates from variation in strategies within the domain of proportional reasoning, providing both a new analytical approach and novel theoretical insight into how best to interpret behavior in these tasks.

Children show difficulty with proportional reasoning when proportional quantities are presented as discrete entities (e.g., a set of red and blue dots, where the target is the proportion of dots that are red) or are difficult to integrate into a coherent continuous whole (e.g., spatially separated red and blue components). For example, when asked to compare two divided game spinners and decide which has a higher

* Corresponding author at: Rutgers University, New Brunswick, 152 Frelinghuysen Rd, Piscataway, NJ 08854, United States of America.
E-mail address: michelle.hurst@rutgers.edu (M.A. Hurst).

https://doi.org/10.1016/j.cognition.2024.105918
Received 3 May 2023; Received in revised form 5 August 2024; Accepted 8 August 2024
Available online 16 August 2024

probability of a winning outcome, children's responses align with the number of winning segments, not the overall proportions (Hurst and Cordes, 2018; Jeong et al., 2007). Similar error patterns are found in many domains that rely on proportional information, including matching juice mixtures (Boyer et al., 2008), interpreting the quantifier "most" (Hurst and Levine, 2022), and making social judgements based on resource distribution (Hurst et al., 2020). These errors, although pervasive, cannot be attributed to a general difficulty with proportional reasoning because children in these same studies succeed when the quantities are continuous and part of an integrated whole. Instead, it may be that children are using a non-proportional heuristic, such as attending to just one quantity alone (e.g., the number of winning segments in a game spinner, the number of items shared) instead of the proportion (e.g., the relative number of winning segments within the spinner as a whole, the number shared relative to how many they started with; Hurst et al., 2020), but doing so selectively with only some kinds of stimuli (Boyer et al., 2008). Furthermore, this pattern is unlikely the result of children having low knowledge of when to use proportion, because adults, like children, also show evidence of interference from numerical information with discrete non-symbolic displays (Fabbri et al., 2012; Hurst et al., 2021). Additionally, adults prefer to use fractions to describe discrete proportion displays and decimals to describe continuous proportion displays, suggesting a conceptual distinction between number-based and area-based proportion (DeWolf et al., 2015).

Notably, one benefit of formalizing strategies is that we can quantitatively compare our different-strategy account with a plausible alternative explanation. It may be that people are not using different strategies, but instead are implementing the *same* strategy with different levels of accuracy across formats. For example, people's behavior on symbolic and non-symbolic magnitude comparison tasks is impacted by the values of the proportional magnitudes (Hurst and Cordes, 2016; Kalra et al., 2020; Park et al., 2020). Thus, it may be that people are trying to use proportion strategies, but it is harder to encode the proportion magnitude from discrete values than continuous values, leading to more noise in our mental representations and therefore more errors with discrete stimuli than continuous stimuli. We incorporate multiple models of proportion comparison strategies that incorporate noise and precision as parameters. If people are trying to use the same strategy, but have different levels of noise or precision, then we would expect to see a similar pattern of model fit, but differences in the estimated parameters for a given strategy.

Across three experiments, we compare strategy use on proportion tasks with different stimuli (Fig. 1). In each experiment, we broadly characterize the stimuli as continuous (i.e., proportion based on area) or discrete (i.e., proportion based on discrete countable objects), but it is worth noting that each experiment uses a different context to situate the proportion comparison task and uses different kinds of discrete and continuous stimuli, which also differ in ways other than just the presence of countable information. In Experiment 1, adults were asked to judge which of two continuous area-based game spinners or discrete number-based vending machines had a higher probability of a red outcome (Fig. 1, top row). In Experiment 2 we re-analyzed data from Park et al. (2020) in which preschoolers, 2nd graders, 5th graders, and adults were asked to judge which of two ratios was larger based on the ratio of separated blue and yellow dot clouds, lines, circles, or blobs (Fig. 1, middle row). In Experiment 3, adults were asked to judge which of two stimuli had a higher proportion of a given color based on the colored area of a single shape (pie chart or blob) or based on the number

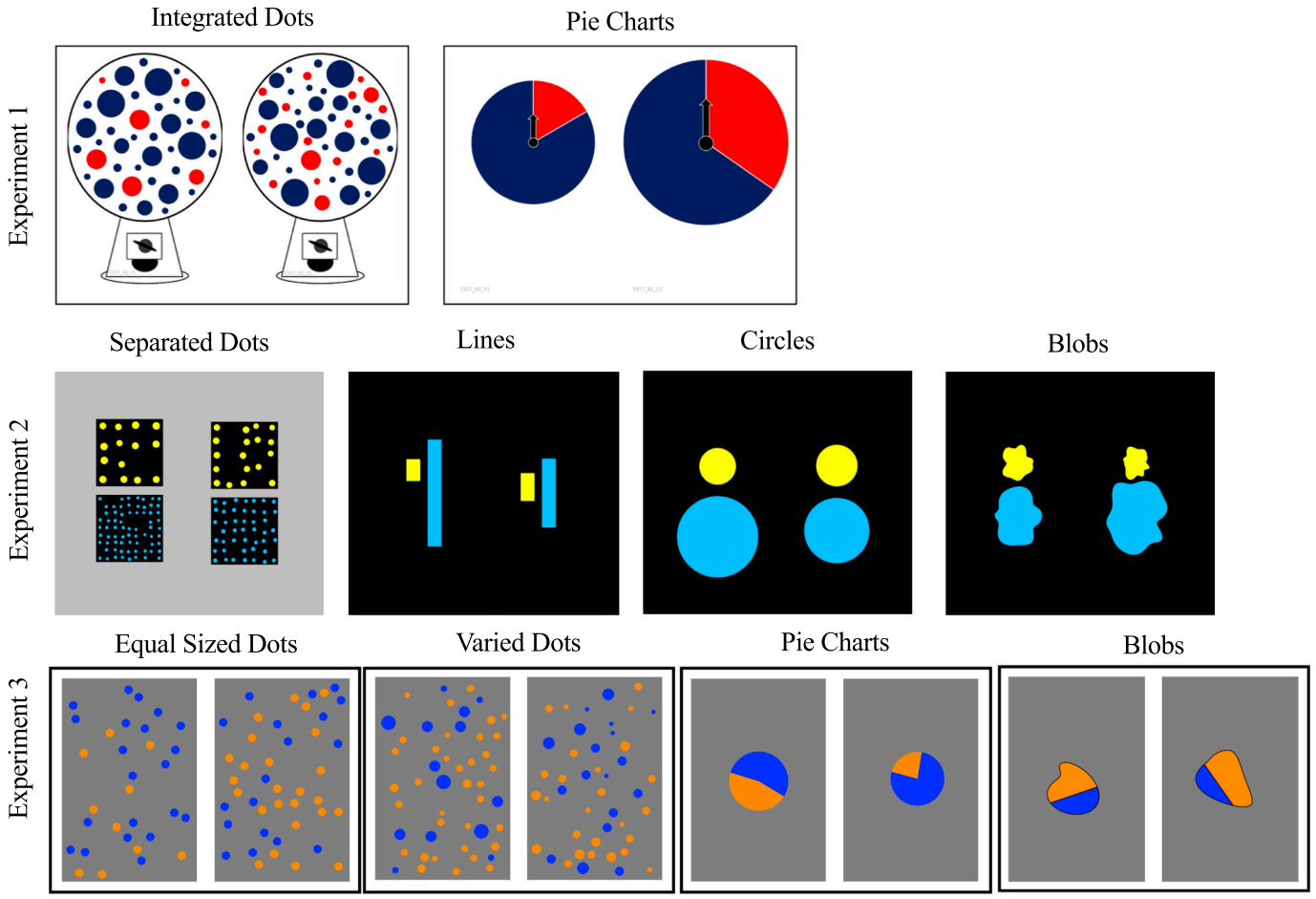

Fig. 1. Examples of stimuli used in Experiment 1 (top row), Experiment 2 (middle row), and Experiment 3 (bottom row).

of colored intermixed dots (with and without controlling for cumulative area; Fig. 1, bottom row). To investigate strategy use, we formalized and tested seven quantitative models, four using proportion strategies and three using non-proportion heuristic strategies. We then compared the Bayesian models of each strategy using model comparison at both the group and individual levels. If it is the case that people use different strategies, as we predict, then we would expect categorically different models to best fit data from different stimuli. In contrast, if people use the same strategies across formats and development but are simply better or worse at using the strategy, then model comparisons should reveal that the same or similar models best fit data from different stimuli, but that the parameters differ.

1.1. Model development

We generated seven models (see Table 1) that differ in the underlying theoretical strategy and the assumed process through which that strategy is carried out. Three models (a, b, c) assume people use only non-proportional information and four models (d, e, f, g) assume people compare proportions but differ in what is encoded and how that information is represented. Each strategy was motivated by prior work with symbolic fractions and non-symbolic proportion (e.g., Boyer et al., 2008; Faulkenberry and Pierce, 2011; Fazio et al., 2016) and formalized based on existing psychophysical models. For all strategies, we model the participant's choice using a Bernoulli function with the probability of selecting the correct response as the probability of success. The probability of selecting the correct response differs across each of the strategies, based on the model assumptions relevant to that strategy.

Table 1
Model parameterizations for each strategy.

	Model Formalization	Parameter Priors
a – Weber Dependent Comparison of Numerators[1]	$\Phi\left[\frac{\lvert n_1 - n_2\rvert}{w\sqrt{n_1^2 + n_2^2}}\right]$	$w \sim exp(1)$ $w > 0$
b – Weber Dependent Comparison of Non-Numerator Components[1]	$\Phi\left[\frac{\lvert r_1 - r_2\rvert}{w\sqrt{{r_1}^2 + {r_2}^2}}\right]$	$w \sim exp(1)$ $w > 0$
c – Weber Dependent Comparison of Denominators[1]	$\Phi\left[\frac{\lvert d_1 - d_2\rvert}{w\sqrt{d_1^2 + d_2^2}}\right]$	$w \sim exp(1)$ $w > 0$
d – Weber Dependent Comparison of Proportions	$\Phi\left[\frac{\lvert p_1 - p_2\rvert}{w\sqrt{{p_1}^2 + {p_2}^2}}\right], where\ p_i = \frac{n_i}{d_i}$	$w \sim exp(1)$ $w > 0$
e – One-Cycle Power Model	$logistic\left(B_1 \left\lvert \frac{n_1^\beta}{r_1^\beta + n_1^\beta} - \frac{n_2^\beta}{r_2^\beta + n_2^\beta} \right\rvert\right)$	$\beta \sim exp(1)$ $B_1 \sim normal\ (0, 3)$
f – Beta Binomial Model	$\Phi\left[\frac{\lvert \mu_1 - \mu_2\rvert}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right], where\ \mu_i = \frac{n_i + \alpha}{(n_i + \alpha) + (r_i + \beta)}$ and $\sigma_i^2 = \frac{(n_i + \alpha)(r_i + \beta)}{(n_i + \alpha + r_i + \beta)^2(n_i + \alpha + r_i + \beta + 1)}$	$\alpha, \beta \sim exp(1)$ $\alpha, \beta > 0$
g – Difference Between Proportions	$logistic(B_1\ \lvert p_1 - p_2\rvert), where\ p_i = \frac{n_i}{d_i}$	$B_1 \sim normal\ (0, 3)$

Note: numerators (n), denominators (d), and non-numerator components (r = d – n) are set by the trial stimuli and subscripts 1 and 2 refer to each of the two stimuli in the comparison; Φ is the cumulative normal distribution.

[1] The equation presented here is the probability correct when the relevant quantity is congruent with the overall proportion (i.e., the larger proportion also has the larger numerator, smaller non-numerator component, or larger denominator). When the relevant component and the overall proportion are incongruent, the complementary probability (i.e., 1 – the equation) is used instead.

For the non-proportion strategies, we assume that people rely on comparing only a single value or quantity across the two stimuli as a heuristic, with each of the three strategies differing in terms of which component people use: numerator (model a), non-numerator component (model b), and denominator (model c). In symbolic fractions, the numerator and denominator refer to the quantities represented above and below the fraction line, respectively (i.e., a/b, a = numerator and b = denominator) and we are defining the non-numerator component as the "left-over" non-numerator amount (i.e., b – a). In general, the numerator represents the amount to be taken out of the denominator or to be divided by the denominator and the denominator represents the whole for the numerator to be taken out of or in reference to. In each experiment, which visible quantity should be treated as the numerator and what is represented as the denominator is defined by the procedure and instructions the participants were given (see each separate Procedure section). In all experiments, we defined the numerator and denominator heuristic strategies as selecting the option with the *larger* component (i.e., larger numerator or larger denominator, respectively), regardless of the other components. For the non-numerator component (i.e., the left-over amount, model b) we defined the strategy as selecting the *smaller* component, with the rationale that participants may be trying to minimize this component, either because it is the unwanted amount or as a strategy to minimize the difference between the numerator and the total (Obersteiner et al., 2022).

For each model, regardless of which component is being compared, we assume that people use an approximate magnitude system for comparing the two quantities. The approximate magnitude system is dependent on Weber's law, with each quantity represented as a normal distribution centered on the true quantity with scalar variability (e.g., Gallistel and Gelman, 1992; Meck and Church, 1983). The true quantities were based on the number of items for discrete sets and visual area (using an arbitrary unit, as described in the Method) for continuous stimuli. The probability of selecting the target option on each trial is modeled as the difference between the two normal distributions using a cumulative normal distribution centered on the absolute value of the difference between the two quantities and standard deviation as the sum of the squared quantities scaled by a constant parameter, commonly referred to as the Weber fraction (w in Table 1), modeled with an exponential prior (Piantadosi, 2016).

Using these assumptions, we mathematically modeled the probability of getting a given trial correct (i.e., selecting the largest proportion that is the target color). For these non-proportion strategies, however, the probability of getting the trial correct depends on the congruency between the quantity being used as a heuristic and the overall proportion. For example, on a trial of 2/3 versus 4/9 the numerator and proportion are incongruent because someone who uses a numerator only strategy would (incorrectly) select 4/9 as being larger (i.e., 4 > 2). In contrast, on a trial of 2/3 versus 3/4, someone who uses a numerator only strategy would select 3/4, which is also the largest proportion. Thus, on incongruent trials the probability of selecting the larger numerator is the complementary probability to selecting the larger proportion. Thus, the equations presented in Table 1 are the probability of selecting the larger proportion (i.e., the correct response) when the larger proportion also has the larger numerator, larger denominator, or smaller non-numerator component (i.e., congruent trials). When the larger proportion has the smaller numerator, smaller denominator, or larger non-numerator component the equation in Table 1 describes the probability of selecting the *smaller* proportion and therefore the complementary probability (i.e., one minus the equation) was used instead.

For the proportion strategies, we use four models that are theoretically similar but differ conceptually and/or mathematically from each other. Specifically, all four models assume participants compare the two proportions, rather than comparing only absolute sub-components of the quantities. But, they are conceptually different in terms of *what* information is encoded in order to represent the proportion (e.g., the numerator and non-numerator component separately represented

versus the relational proportion as a single value), the mathematical and psychophysical assumptions about *how* that information is represented (e.g., following Weber's law vs. Stevens' law), and the way the comparison between the two proportions is mathematically modeled. In the current manuscript, we include all four models because there is not consensus in the literature of which best describes behavior on proportion comparison tasks, either theoretically or mathematically. We also note that it is difficult to adjudicate between these models, because they make very similar predictions.

In the Weber-dependent comparison strategy (model d), we assume that people represent the proportion as a single value using a magnitude representation system with the same properties as the absolute magnitude systems described for the non-proportion strategies. There is evidence of ratio-dependent responding for symbolic and non-symbolic fractions, which is typically associated with this model of magnitude representation (Hurst and Cordes, 2016; Kalra et al., 2020; Park et al., 2020). Specifically, we assume that the proportions are represented as normal distributions centered on the true value with scalar variability. We then modeled the probability of selecting the correct (i.e., larger) proportion as a cumulative normal distribution with mean as the difference between the two proportions and standard deviation as the sum of the squared proportions scaled by a constant parameter (*w* in Table 1) with an exponential prior.

The one-cycle power model (model e) is based on substantial prior work suggesting that the numerator and non-numerator (i.e., "left over") components are each encoded separately in accordance with Stevens' power law and then combined to create a proportion (Hollands and Dyre, 2000; Spence, 1990). Stevens' power law describes the relation between the actual magnitude of a stimulus and the perceived magnitude of the same stimulus as a power function (e.g., Stevens, 1957). In Hollands and Dyre's one-cycle power model, each of the two subcomponents of the proportion (i.e., the numerator and the leftover difference between the denominator and the numerator) are represented as separate absolute quantities, with the perceived quantity modeled as a power of the true quantities (i.e., using the psychophysics of Stevens' power law), using a parameter β with an exponential prior. These components are then combined to compute the proportion as perceived numerator / (perceived numerator + perceived non-numerator component). Given that the one-cycle power model has been developed only as a model of proportion encoding based on estimation data, it does not provide a theoretically informed way of modeling the comparison process (i.e., deciding which of two proportions is larger, after they are represented). Thus, after each proportion is computed following the one-cycle power model, we modeled the probability of correctly comparing two proportions using a simple logistic function on the difference between the two computed values, with a normal(0,3) prior on the slope parameter B_1. We will also note that the one-cycle power model described by Hollands and Dyre (2000) includes a general form that can incorporate multiple cycles caused by using additional benchmarks. Given that the stimuli used in the current study did not provide any benchmarks, we only use the simple one-cycle form.

The Beta Binomial model (model f) is a foundational model in Bayesian probability when modeling unknown proportions. As with the one-cycle power model (e), this model assumes that both the numerator and non-numerator leftover components are encoded separately. However, this model differs in the mathematical instantiation of both how each component is represented and how they are compared. Specifically, we assumed both the numerator and non-numerator subcomponent were encoded with adjustment parameters α and β, each with an exponential prior. The components were then combined to compute the proportion (adjusted numerator/(adjusted numerator + adjusted non-numerator subcomponent)), with a beta distribution based on the α and β parameters. To compare the two proportions, we calculated the cumulative normal distribution based on the difference between the beta distributions using a normal approximation with mean and standard deviation as described in Table 1 (Cook, 2012). Thus, this approach models the comparison process in a way that is similar to how the comparison process is instantiated in Weber dependent models described above.

Finally, we also report a simplified model of comparison, without theoretically informed assumptions about how the proportions are encoded or compared. In this model (g), each proportion is represented veridically (e.g., ¾ is represented as 0.75) and we model the probability of correctly comparing the two proportions as a logistic function on the absolute difference between the two proportions, scaled by a slope parameter with a normal(0,3) prior. Given that there is not a clear consensus of the psychophysical model that is best for modeling proportion comparison tasks, we include this simplified model as a baseline.

To test how well our stimulus sets were able to distinguish these models, we first generated simulated datasets that corresponded to the use of a given strategy (with some preset parameter values) and then applied our analysis approach to each simulated dataset. If each model was fully discriminable from the others, then we should find that the strategy used to generate a given dataset fit the data substantially better than all other models and the inferred parameters for that strategy should also correspond to the parameter values used to generate the dataset. For the stimuli used in Experiments 1 and 2, we found that the four proportion models (d, e, f, g) made very similar predictions and were rarely fully discriminable. In general, the heuristic models were more separable, but for some extreme parameter values (i.e., relatively small or large values of *w*), they became less discriminable. However, we retained all seven strategies given their theoretical interest and, for the proportion models specifically, because it is unclear which would be the most appropriate model to retain to represent the proportion strategies. Importantly, this does suggest that any null differences (i.e., overlapping predictions and model fitting metrics) among the proportion models (d, e, f, g) should be interpreted as not being able to distinguish between them with these stimuli.

2. Experiment 1

2.1. Method

2.1.1. Participants

One hundred and nine adults (M_{age} = 26 years, Range: 18 to 63 years; 76 women, 33 men) are included in the analyses. Adults participated entirely online and were recruited from participant databases that included university students and community members. Adults received course credit or $5. Eight participants completed the study twice and only their first response is used in the analyses. Prior to completing this task, participants completed a separate experiment investigating their use of quantitative information when making social evaluations of others (Hurst et al., 2020) and the sample size was chosen to provide adequate power for this other study.

2.1.2. Stimuli and materials

Adults completed 80 trials across two blocks, with each block presenting stimuli with a different format (Fig. 1). The order of the two blocks was counterbalanced and the order of trials within a block was randomized. Each block contained 40 unique trials, 10 from each of four ratio bins to ensure variability in closeness between the two proportions (larger proportion of red/smaller proportion of red): 1.06, 1.25, 1.5, and 2.

Discrete stimuli were presented as red and blue dots intermixed within a dispenser. The number of dots of a single color ranged from 6 to 41 and the total number of dots ranged from 14 to 50. The sizes of the dots within a stimulus varied so that the red:blue ratio in terms of surface area did not correspond to the red:blue ratio in terms of number. Proportion comparisons were selected so that the stimulus with the higher number of red items also had the higher proportion of red items on half the trials. Continuous stimuli were presented as circular spinners with a red portion and a blue portion, and a black arrow extending

upward from the center of the circle along a red-blue boundary. The same proportion magnitudes used in the discrete trials were used in the continuous trials. However, on each trial the sizes of the two spinners differed so that the stimulus with the higher red area also had the higher proportion of red on approximately half the trials (see Fig. 2 for examples of these different stimulus features).

This means that the corresponding trial did not have the same absolute component values across format. For example, a discrete comparison of 2/3 vs. 4/8 might correspond to a continuous comparison of 60 pixels / 90 pixels vs. 40 pixels / 80 pixels. Given that the continuous pie charts can be quantified using any unit for the area of each color (e.g., 1 cm / 10 cm = 10 mm / 100 mm), for analysis we used an arbitrary unit that ensured the continuous values were in a similar range as the counts used in the discrete trials (total dots ranged from 16 to 50 dots; total continuous amount ranged from 16 to 62 units).

To ensure all strategies were equally available in the stimulus set, we removed trials that prevented the use of some strategy. In the discrete stimuli, one trial included two stimuli that had the same non-numerator subcomponent (i.e., difference between the numerator and denominator; 18/35 vs. 14/31), making the model b strategy useless. In the continuous stimuli, after scaling and rounding the values to be better aligned with the discrete counts, one trial included stimuli with the same numerator value, making the numerator comparison strategy (model a) useless. Each of the two trials was excluded from their respective format type for all modeling analyses, resulting in 39 useable trials for each format.

2.1.3. Procedure

Adults completed both the discrete and continuous blocks and were randomly assigned to complete the discrete block first ($n = 54$) or the continuous block first ($n = 55$). The procedure within each block was identical and all that differed was the format of the stimuli (see Fig. 3). Written instructions were provided on the screen prior to each block. The instructions introduced participants to the color machines (dispensers in the discrete block and spinners in the continuous block) and instructed participants to select the color machine that had a higher probability of resulting in red. Participants responded by pressing the right or left arrow key for the right or left stimulus, respectively, and were told to respond as quickly as possible. Stimuli remained visible until a response was selected. Adults participated online through Gorilla (www.gorilla.sc; Anwyl-Irvine et al., 2020). We did not restrict the type of device participants used to participate in the study.

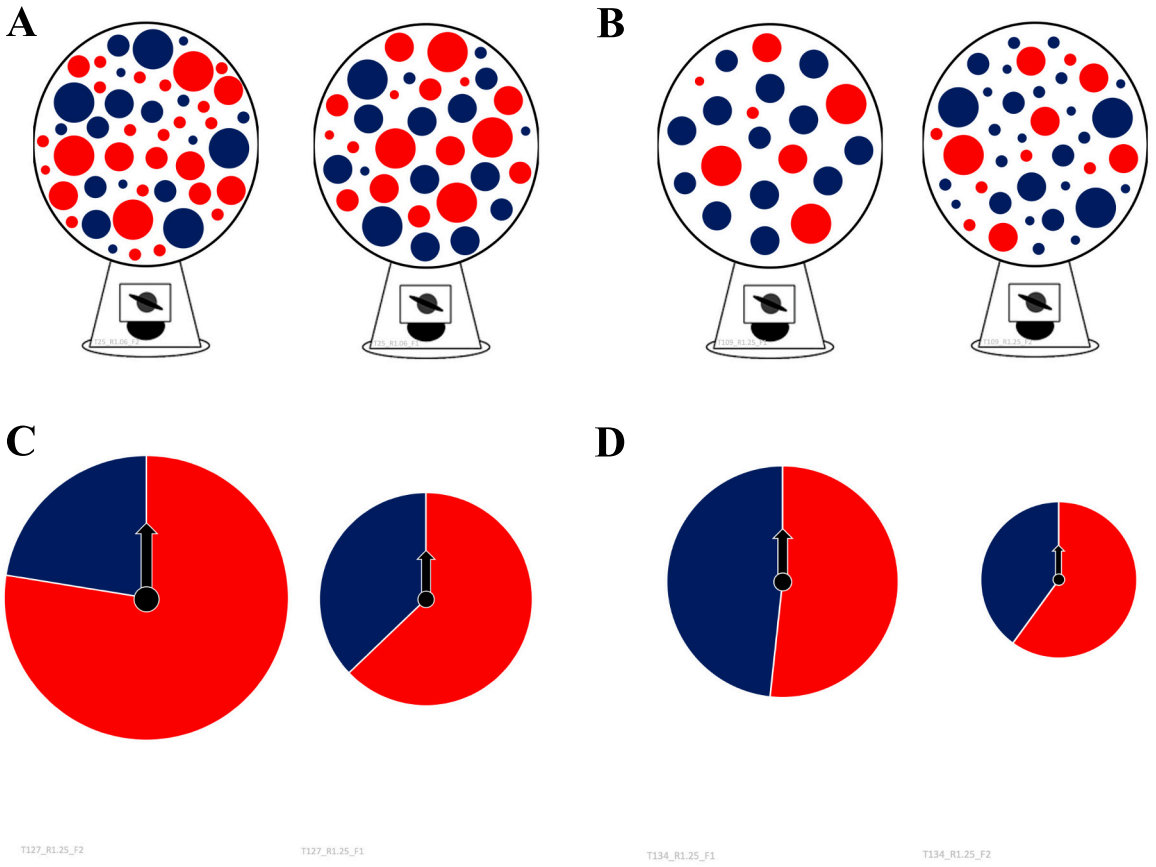

Fig. 2. Examples of different trial types from Experiment 1. A: A comparison of 28 red out of 47 total dots (~ 60 % red; right stimulus) versus 17 red out of 33 total dots (~52 % red; left stimulus). This is a numerator congruent trial because the larger proportion of red dots (right) also has the highest number of red dots (right). Dot size varies so that in each stimulus the cumulative area of blue and red is about the same. B: A comparison of 7 red out of 20 total dots (35 % red; left stimulus) versus 12 red out of 40 total dots (30 % red; right stimulus). This is a numerator incongruent trial because the larger proportion of red dots (left) has a fewer number of red dots. Dot size varies so that in each stimulus the cumulative area of blue and red is about the same. C: A comparison of a spinner with 77 % red (scaled to have ~48 arbitrary units of red, see text) versus 63 % red (scaled to have ~22 arbitrary units of red, see text). This is a numerator congruent trial because the larger proportion of red (left) also has the larger red area (left). D: A comparison of a spinner with 52 % red (scaled to have ~18 arbitrary units of red, see text) versus 60 % red (scaled to have ~9 arbitrary units of red, see text). This is a numerator incongruent trial because the larger proportion of red (right) has a smaller red area. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

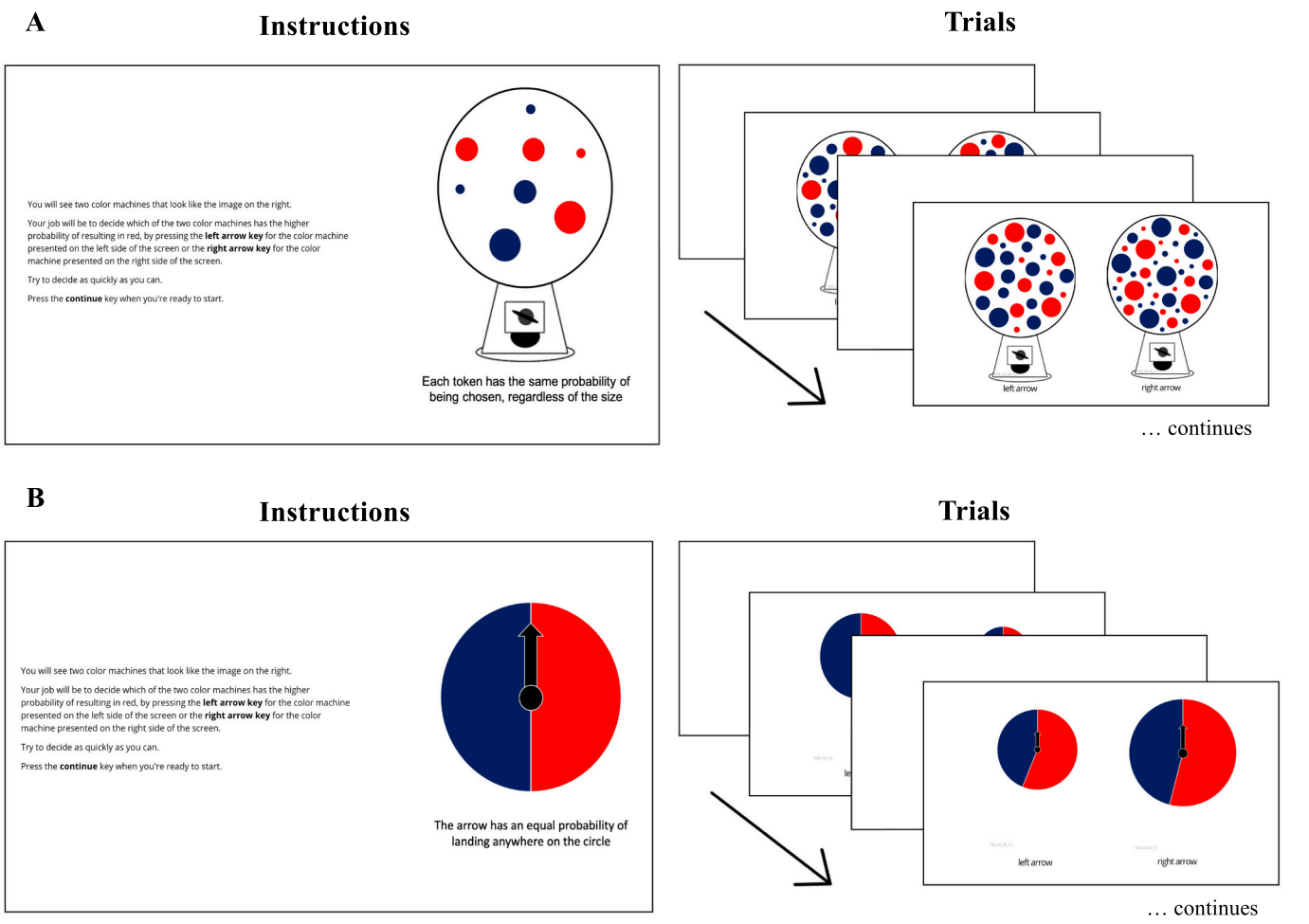

Fig. 3. Schematic of the paradigm used in Experiment 1, separated across the discrete block with integrated dots (Panel A) and the continuous block with pie charts (Panel B). The instructions were shown at the beginning of the block, followed by the trials alternating between the stimuli, with a reminder of the keys to press (visible until response selected) and a blank screen between trials (500 ms).

2.1.4. Analytical details

All data analyses and visualizations were done with R version 4.0.2 (R Core Team, 2020) and R Studio version 2022.07.1 (R Studio Team, 2016). Data cleaning and organization used the *tidyverse* 1.3.0 (Wickham, 2017), models were written and fit using *rstan* version 2.21.3 (Stan Development Team, 2020), model comparison was completed using *loo* version 2.5.1 (Vehtari et al., 2022), and plots were created using *ggplot2* version 3.3.6 (Wickham, 2016).

2.1.5. Group-level models

Each model in Table 1 was separately fit to the data from each format. Parameters of each model were inferred using a No-U-Turn sampler (Hoffman and Gelman, 2014) sampling four chains each with 5000 iterations (except for the Beta Binomial Model, which had 10, 000 iterations to consistently converge), with 50 % as warm up. Model diagnostics showed no divergent transitions, appropriately large effective sample sizes (minimum = 1028), and all $\widehat{R} = 1$, suggesting the models did converge.

Point-wise log likelihoods were generated during model fitting and were extracted for computing both LOO and WAIC model comparison metrics. We also use the ELPD (expected log pointwise predictive density) as a measure of predictive accuracy (Vehtari et al., 2017). These values are not independently meaningful but can be compared for different models when estimated on the same data, with lower WAIC or LOO values and higher ELPDs suggesting a better fitting model. Note that this means that only the ordinal pattern and not the specific values can be compared across formats, age groups, and experiments, because the models are estimated using different datasets or subsets of the datasets.

2.1.6. Individual models

To test whether the overall pattern was consistent across individuals, we used the same modeling approach on each individual participant's data separately. These analyses rely on smaller datasets, resulting in issues with model fitting for some participants and the LOO and WAIC estimates using importance sampling were more susceptible to outlying trials (see Supplemental Materials C for details).

2.1.7. Transparency and openness

The sample size and basic design was pre-registered (https://aspredicted.org/bx523.pdf). An analysis plan was also pre-registered; however, the analysis plan was based on a traditional frequentist approach to investigating adults' behavior when comparing discrete versus continuous proportion (the pre-registered analyses can be found in Hurst and Piantadosi (2022) and at https://osf.io/bescn. In the current manuscript, we re-use this data (Hurst, 2022) with the different, and not pre-

registered, goal of quantitatively comparing adults' strategy use with Bayesian models of behavior. All models[1] and decision criteria are transparently reported throughout the manuscript. Data and analysis scripts (https://osf.io/2rtdq) and materials (https://osf.io/bescn) are publicly available. Participants provided informed consent and the study was approved by the University of Chicago Institutional Review Board IRB 17-1599, "Relational Math Reasoning".

2.2. Results and discussion

Data from each format was separately fitted with each of the models in Table 1 and compared using information criteria and expected log predictive density (ELPD) model comparison estimates calculated using the Widely Applicable Information Criterion (WAIC) and leave-one-out procedure (LOO). Parameter estimates from each model and model comparison metrics are reported in Supplemental Materials A and B, respectively. The differences between ELPD of the best fitting model and all other models are presented in Fig. 4 and throughout we will interpret this difference ($ELPD_{diff}$) relative to the estimated standard error of the difference (SE_{diff}). For ease of reporting, we define the difference ratio as $D = ELPD_{diff} / SE_{diff}$, so that D can be interpreted as the number of standard errors between the model fit metrics of the best fitting model and the comparison model. Additionally, for a subset of the models, trial level model predictions of accuracy, empirical accuracy (i.e., proportion correct), and the ratio between the two numerators (larger numerator/smaller numerator) or the two proportions (larger proportion/smaller proportion) are provided in Fig. 5, as well as R^2 estimates from linear regressions predicting empirical accuracy from model predictions at the trial level. Although we cannot directly assess absolute model fit, the R^2 estimates, combined with qualitative visual inspection of the relation between empirical data and model predictions, can provide some insight.

For pie chart stimuli, the proportion models better fit the data than the non-proportional heuristic models. Within the proportion models, the one-cycle power model (e) best fit the data, with most of the other proportion models being indistinguishable (e vs. d, g: $Ds < 3$), except for the beta binomial model (f), which was moderately worse ($D = 6.1$). The non-proportion heuristic models (a, b, c) all fit the data substantially worse than the best fitting model ($Ds > 16$). For the integrated dot stimuli, in contrast, the numerator comparison model (a) best fit the data, followed by the beta binomial proportion model (f) ($D = 4.9$), with all other models worse fitting ($Ds > 7$).

All model parameters are reported in Supplemental Materials A and here we briefly discuss parameter estimates from the most relevant models. With integrated dot stimuli, the Weber estimate from the numerator comparison model (a) ($w = 0.66$) is higher than typically reported for number comparisons in other studies, which tend to be around 0.1 for educated adults in the USA (e.g., Odic et al., 2013). With pie chart stimuli, the proportion weber estimate (d) ($w = 0.25$) is lower than that found for numerator comparisons in the current dataset but is still fairly high for adults. It must also be noted, however, that this value should be interpreted on a bounded proportion scale, making it difficult to compare to absolute judgements. That is, unlike numerical representations, which can scale infinitely, the bounded nature of proportion (i.e., values are between 0 and 1) impacts the estimate of the noise parameter. For the one-cycle power model (e), the beta parameter estimate (beta = 1.36) is surprising because it is greater than one. Typical beta estimates for area based or number-based proportions are less than one, which accounts for the typical bias seen in estimation data of overestimation of values under half and underestimation of values above half (Hollands and Dyre, 2000). Beta estimates greater than one correspond to a reversal of this bias pattern, which is not typically found for number or area (although, there are some exceptions, see Shuford, 1961).

Finally, at the individual participant level, there was substantial variability within and between people, with smaller differences between ELPD estimates making it more difficult to distinguish the strategies (i.e., there was rarely a single best fitting model with $D > 3$; see Supplemental Materials C for details). Overall, however, the ordinal relation between model fit estimates (i.e., regardless of similarity between models) shows the same general pattern: on pie chart data, for most adults a proportion model (d, e, f, g) was numerically the best fitting model (98/109), but on the integrated dots data, adults were more evenly split between the numerator model (a) (47/109, which was the modal best fitting model) and a proportion model (d, e, f, g) (44/109, although the specific proportion model varied substantially). However, when considering the similarities between model fits, for most adults both the numerator model (a) and at least one proportion model (d, e, f, g) similarly fit the integrated dots proportion data ($Ds < 3$, 87/109). For a smaller group, none of the proportion models (d, e, f, g) fit as well as the numerator model (a) ($Ds > 3$, 11/109) or the numerator model (a) fit worse than a proportion model ($Ds > 3$, 11/109). For pie chart data, again for most adults both the numerator model (a) and a proportion model (d, e, f, g) similarly fit the continuous proportion data ($Ds < 3$, 63/109), but, in contrast to integrated dots stimuli, there was another sizeable group where the numerator model (a) fit worse than a proportion model ($Ds > 3$, 45/109) and for only a single participant did the numerator model (a) fit better than all proportion models (d, e, f, g).

3. Experiment 2

3.1. Method

3.1.1. Participants

In our re-analysis, we analyzed data from 36 preschoolers (4–5-year-olds), 28 2nd graders (approximately 6–7-year-olds), 29 5th graders (approximately 9–10-year-olds), and 32 adults (approximately 19–20-year-olds). This includes more data than the analyzed sample in Park et al. (2020) because we did not exclude individual trials based on reaction times or entire participants based on their overall accuracy. Given that the goal of Park et al. (2020) was to determine precision differences across formats and age groups, excluding trials and participants that are likely guessing is a sensible approach to isolate data that more likely capture people's actual ratio comparison ability. However, for our purposes of strategy discovery, these trials and participants are important for analyzing the full set of possible behaviors. More details about the sample can be found in Park et al. (2020).

3.1.2. Stimuli and materials

On each trial of the ratio comparison task, two stimuli were presented, one each on the left and right side of the screen. The ratio stimuli were presented as yellow and blue line lengths, circles, blobs, or sets of dots, with each format in a separate block (Fig. 1). In each stimulus, the ratio ranged from 1:5 (i.e., 0.2) to 4:5 (i.e., 0.8). The ratio of ratios used on each trial came from one of five ratio bins: 3:1, 2:1, 2:3, 3:4, and 5:6. The stimuli remained visible for 1500 ms for adults, 6000 ms for 2nd and 5th grade children, or until a response was selected for preschoolers.

[1] Previous versions of our analysis also included additional models: (1) a comparison of the sums of the numerator and denominator, (2) a modified one-cycle-power model of the numerator and denominator rather than the numerator and non-numerator component, and (3) area-based comparisons using perceived area estimates based on Yousif and Keil (2019). The sum model was included because it has been hypothesized as a possible incorrect strategy used with symbolic numbers, but the structure of visually presented proportional information makes it a non-sensical strategy in this case. The modified one-cycle-power model and the additive-area models were included as possible alternatives to the typical one-cycle power model and for modeling perceived area. However, neither were as theoretically justified and did not fit the data meaningfully differently than the models included here. For clarity, and given the low theoretical motivation for either model, we have opted to exclude them from the manuscript, but report them here for transparency.

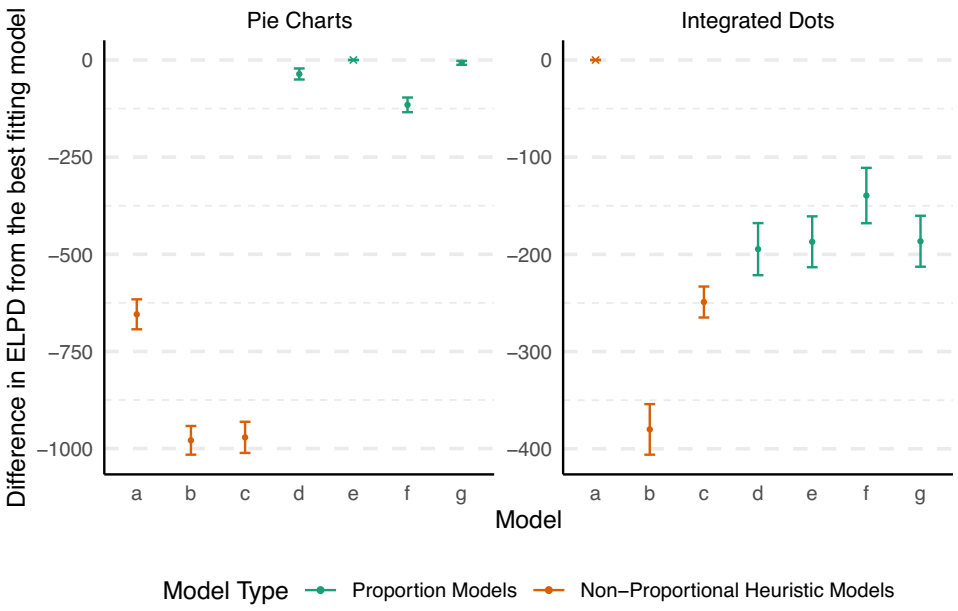

Fig. 4. Differences in Expected Log Predictive Densities (ELPD) from the best fitting model, separately for the continuous pie charts (left panel) and discrete dots (right panel). The best fitting model is represented as an x and has a value of zero. Error bars are estimated standard errors of the difference from this model. Non-proportion heuristic models (a, b, c) are in orange and proportion models (d, e, f, g) are in green (see Table 1 for details of each lettered model). a = Weber-Dependent Comparison of Numerators; b = Weber-Dependent Comparison of Non-Numerator Components; c = Weber-Dependent Comparison of Denominators; d = Weber Dependent Comparison of Proportions; e = One-Cycle Power Model of Proportions; f = Beta Binomial Model of Proportions; g = Logistic Model on the Difference Between Proportions.

During the task, children completed 40 trials per format and took a break every 10 trials, while adults completed 60 trials per format and took a break every 20 trials.

Dots were presented as two sets of dots inside black squares that were spatially separated and aligned vertically with the yellow set on top and blue set on the bottom. Area was controlled such that the cumulative area was equated across the two ratios. The number of yellow dots (i.e., numerator) ranged from 12 to 92 and the number of blue dots (i.e., denominator; see Procedure) ranged from 20 to 120. Lines were presented next to each other, with the yellow on the left and blue on the right. The blue lines were vertically centered, and the yellow line was given random vertical jitter to be unaligned with either the top, bottom, or center of the blue line. The yellow line ranged from 35 to 228 pixels long and the blue line ranged from 50 to 304 pixels long. Circles were presented vertically, with the yellow on the top and blue on the bottom. The circles were horizontally centered relative to each other. The yellow circle ranged from 1602 to 11, 540 square pixels and the blue circle ranged from 2289 to 15, 386 square pixels. Blobs were presented vertically, with the yellow on the top and blue on the bottom. The blobs were horizontally centered relative to each other. The yellow blobs ranged from 1467 to 3518 square pixels and the blue blobs ranged from 4891 to 29, 348 square pixels.

3.1.3. Procedure

In addition to the ratio comparison task, participants completed an absolute magnitude comparison task and an inhibitory control task. These tasks were administered before the ratio comparison task, in the same session for adults or in a different previous session for children. However, performance on these tasks is not relevant to the current manuscript and so we do not discuss them further (see Park et al., 2020 for details).

On the ratio comparison task, the line and circle tasks were always completed first followed by the blob and dot tasks, with the order of each task within these pairs counterbalanced. Prior to each format, children received format-specific instruction on ratios. This instruction included an introduction to ratios and described how ratios get larger as the components (i.e., the blue and yellow pieces) become more similar. Notably, the task instructions describe the process of comparing *ratios*, not proportions (e.g., there is no part/whole structure). To map these stimuli onto our terminology and the way we have operationalized the strategies, we define the smaller component (yellow) as the numerator and the larger component (blue) as the denominator. Additionally, all participants completed 12 practice trials per format, which were repeated if children's accuracy was lower than 50 % on the first set of practice.

3.1.4. Analytical details

Data was analyzed as described for Experiment 1. Again, model diagnostics showed no divergent transitions, appropriately large effective sample sizes (minimum = 1470), and all $\widehat{R} = 1$, suggesting the models did converge.

3.1.5. Transparency and openness

This is a re-analysis of previously published data, not collected by the current authors. Data was retrieved in a CSV file from the original authors (Park et al., 2020). Data, materials, and additional information about the original paper can be found in their publicly available repository (linked in Park et al., 2020). Data and analysis scripts used for the current manuscript can be found in our OSF repository (https://osf.io/2rtdq). The original data was collected and disseminated according to the original authors IRB protocols. Re-use of data from Park et al., (2020) was deemed not human subjects research by the University of Chicago IRB.

3.2. Results and discussion

In Experiment 2, we used the same analytical strategy but now have data from four stimuli formats (dots, lines, circles, blobs) and four age groups (preschoolers, 2nd graders, 5th graders, adults). Point estimates of the model parameters and model comparison metrics are presented in Supplemental Materials A and B, respectively. Differences in ELPD values, as described in Experiment 1, are presented in Fig. 6 and comparisons of model predictions versus empirical accuracy are presented in

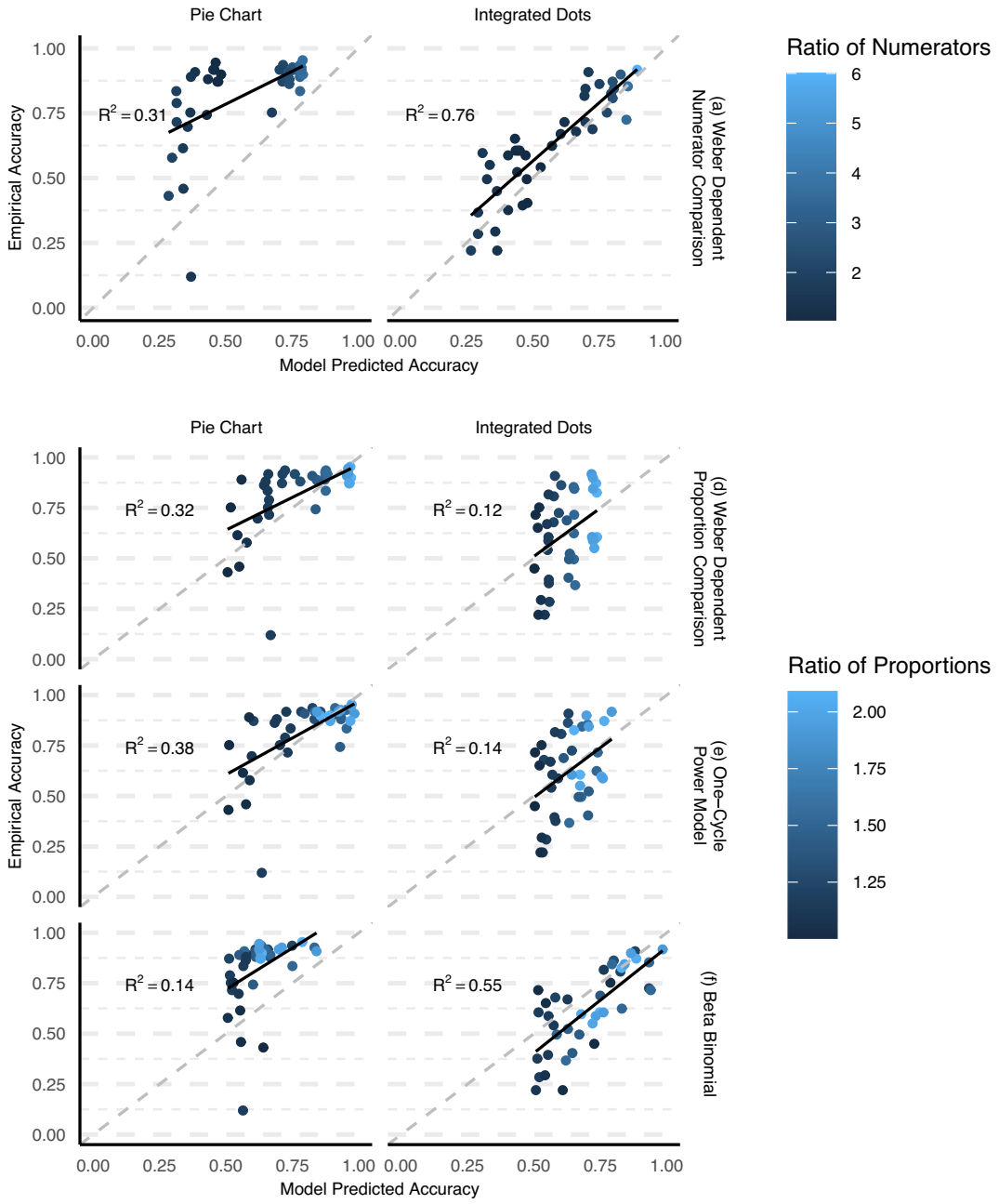

Fig. 5. Model predicted probability correct (x-axis) and empirical accuracy (proportion correct; y-axis), colored as a function of the ratio between the numerators (larger numerator/smaller numerator) for the numerator comparison model (a) and between the proportions (larger proportion/smaller proportion) for the proportion models. Model estimates were based on models using the best fitting parameter point-estimates. Each point is summary data across participants for a single trial. The dotted diagonal line is along x = y, as an indicator of what perfect agreement would look like. The R^2 estimates were calculated from linear regression models predicting empirical accuracy from model predicted accuracy, summarized at the trial level.

Fig. 7.

For line ratio comparisons, the best fitting model for all age groups was the Weber dependent proportion model (d), which was indistinguishable from most of the other proportion models for all age groups (d vs. e, g: Ds < 4), except the Beta Binomial model (f), which varied by age group (preschool D = 5.1; 2nd grade D = 4.6; 5th grade and adults D $<$ 3). In contrast, the numerator comparison model (a) was worse fitting for all age groups and the difference was much larger for adults than for children (preschool D = 5.1; 2nd grade D = 6.2; 5th grade D = 6.4; Adults grade D = 15.6). All other non-proportional heuristic models (b, c) were substantially worse fitting (all Ds > 5), except the comparison of the non-numerator component (b) for 2nd graders, which was only moderately worse fitting, D = 4.4.

For blob ratio comparisons, the best fitting model for all age groups was the Weber dependent proportion model (d). For preschoolers and 5th graders, this model was similar to the other proportion models (d vs. e, g; all Ds < 4), except the Beta Binomial model (f) which was moderately worse fitting ($4 <$ D < 6). For 2nd graders and adults, all other proportion models were moderately worse fitting (d vs. e, f, g; $4 <$ D < 7). The pattern was more mixed across development for the numerator comparison model. For preschoolers, the numerator comparison model (a) was indistinguishable from the best fitting model (d vs. a; D = 0.7), for 2nd graders and 5th graders it was moderately worse fitting (2nd grade D = 6.1; 5th grade D = 7.7), and for adults it was substantially worse fitting (D = 18.5). All other non-proportion heuristics were at least moderately worse fitting (all Ds > 5).

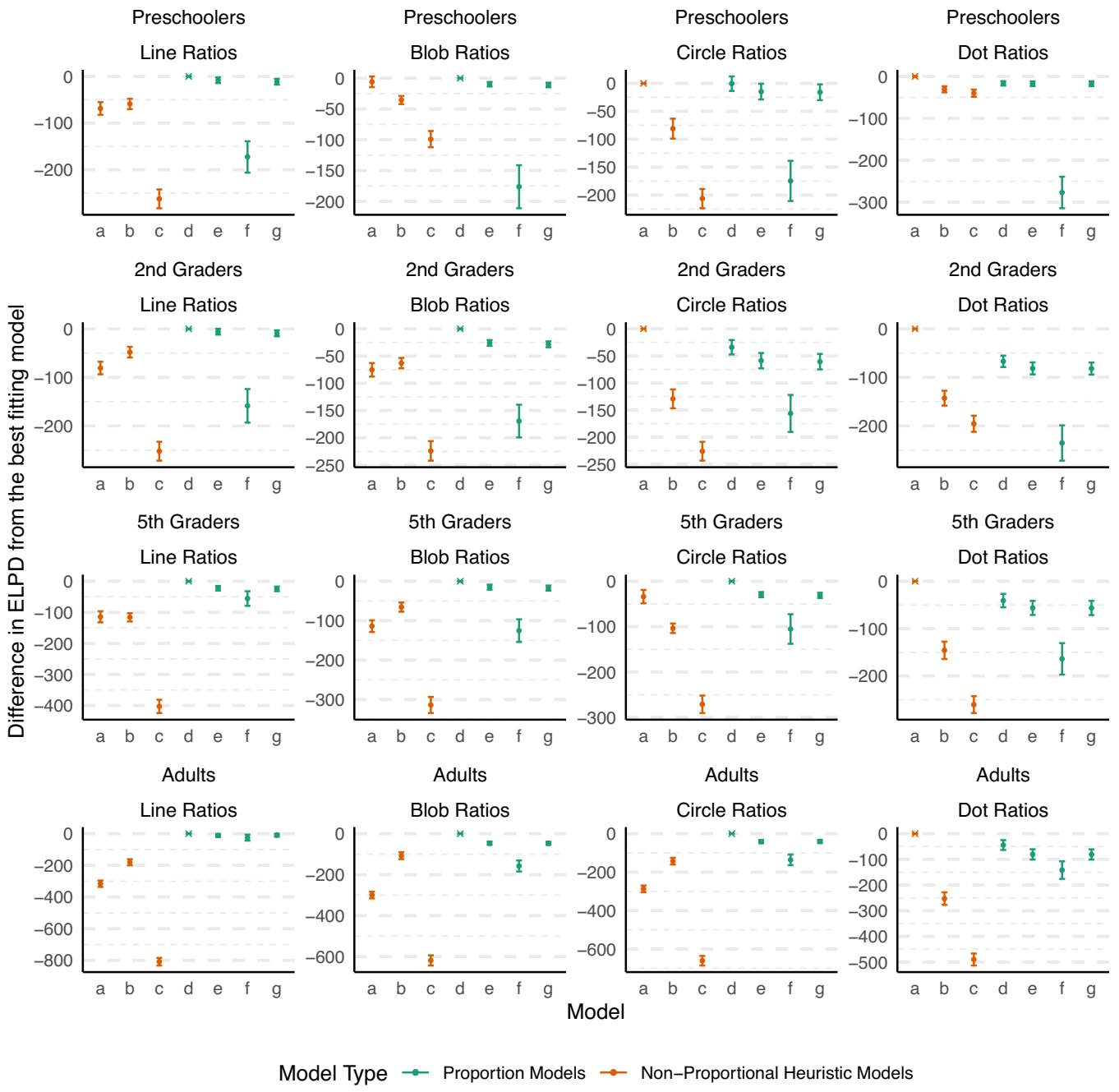

Fig. 6. Differences in Expected Log Predictive Densities (ELPD) from the best fitting model, separately for each format and age group. The best fitting model is represented as an x and error bars are estimated standard errors of the difference from this model. Non-proportion heuristic models (a, b, c, d) are in orange and proportion models (e, f, g, h) are in green (see Table 1 for details of each lettered model). a = Weber-Dependent Comparison of Numerators; b = Weber-Dependent Comparison of Non-Numerator Components; c = Weber-Dependent Comparison of Denominators; d = Weber Dependent Comparison of Proportions; e = One-Cycle Power Model of Proportions; f = Beta Binomial Model of Proportions; g = Logistic Model on the Difference Between Proportions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

For circle ratio comparisons, data from preschool and 2nd grade children was best fit by the numerator comparison model (a), followed by the Weber dependent proportion model (d; $D < 3$). The remaining proportion models were also indistinguishable or at most moderately worse fitting (e, f, g; $D < 5$). Data from 5th graders and adults were best fit by the Weber dependent proportion model (d), which was similar to or only moderately better than the other proportional models (e, f, g; $D < 6$). Only in the adult sample was the best fitting proportion model (d, for adults) substantially better than the numerator comparison model (a) ($D = 16.6$). All other non-proportion heuristic models (b, c) were substantially different from the best fitting model in all age groups ($Ds > 7$), except the non-numerator component model for preschoolers which was only moderately worse fitting than the best fitting model (a, for preschoolers; $D = 4.5$).

For dot ratio comparisons, in all age groups the numerator model (a) was the best fitting model. However, only in 2nd graders were all other models worse fitting ($Ds > 5$). In 5th graders and adults, the proportion models (d, e, f, g) fit similarly to the numerator model (a) ($Ds < 5$) and the other non-proportional heuristic models were substantially worse fitting (a vs. b, c; $Ds > 7$). For preschoolers, most of the models fit similarly with proportional and non-proportional models intermixed (a vs. all; $3 < Ds < 8$).

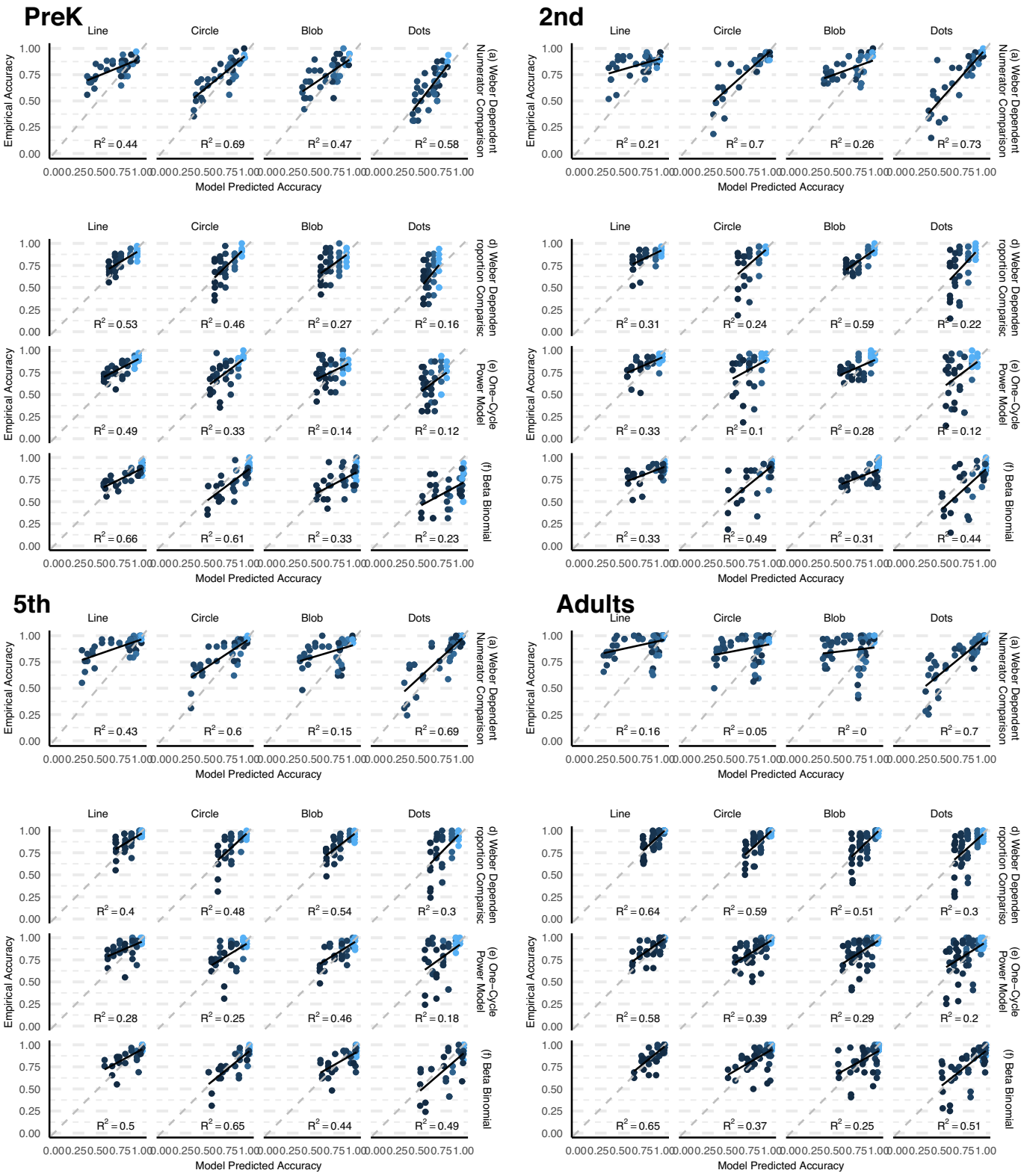

Fig. 7. Model predicted probability correct (x-axis) and empirical accuracy (proportion correct; y-axis) at the trial level from Experiment 2, colored as a function of the ratio between the numerators (larger numerator/smaller numerator) for the numerator comparison model (a) and between the proportions (larger proportion/smaller proportion) for the proportion models (d, e, f). Model estimates are based on models using the parameter point-estimates for each age group and format separately. Each point is summary data across participants for a single trial. The dotted diagonal line is along x = y, as an indicator of what perfect agreement would look like. The R^2 estimates were calculated from linear regression models predicting empirical accuracy from model predicted accuracy, summarized at the trial level.

With dot stimuli, the Weber parameter estimate from the numerator comparison model (a) shows the developmental trend we would expect (i.e., decreasing with age) but are higher than typically reported for absolute number comparisons, $w_{\text{PreK}} = 0.67$, $w_{\text{2nd}} = 0.34$, $w_{\text{5th}} = 0.29$, $w_{\text{adults}} = 0.27$. With continuous stimuli (lines, circles, blobs), the proportion Weber estimates (d) also decrease developmentally and are lower for lines than circles or blobs (range: 0.18 to 0.52), consistent with typical developmental patterns and with the pattern found in our results of more consistent proportional reasoning behavior with lines than circles and blobs. For the one-cycle power model (e), the beta parameter estimates are generally below one and around the values typically reported in other studies using an estimation task (Range: 0.68–0.82; Hollands and Dyre, 2000), except for adults whose parameter estimates are around or slightly above one (Range: 0.98–1.35), as was found in Experiment 1.

When looking at each individual separately (see Supplemental Materials C), the same general pattern was found, but again with substantial variability and smaller differences in ELPD between models (i.e., the models were less distinguishable at the individual level). Based on just the ordinal pattern between models, most adults and 5th graders were best fit by a proportion model (d, e, f, g) on the continuous stimuli (Adults: lines 27/32; circles 27/32; blobs 27/32; 5th graders: lines 22/29, circles 18/29, blobs 21/29). Preschool aged children and 2nd graders showed more variation across the three kinds of continuous stimuli. Most of these children were best fit by a proportion model (d, e, f, g) when comparing lines (Preschool 19/27; 2nd grade 19/27) and blobs (Preschool 7/18, 2nd grade 16/27), but there was a mix between numerator (a) and proportion models (d, e, f, g) for circles (Preschool 17/34 numerator vs. 12/34 proportion; 2nd grade 17/27 numerator vs. 8/27 proportion). When comparing dots, preschoolers showed a mix of strategies, without a clear modal preference. Most 2nd and 5th graders were best fit by a numerator model (a) (2nd grade 24/27; 5th grade 19/29), while adults were almost equally split (16/32 numerator vs. 16/32 proportion strategies).

As a final way to explore these patterns, we looked at older children's and adults' tendency to use the same strategy type when comparing lines (as a representative of "continuous" stimuli) and dots (we omit preschool children here because they did not show robust strategy preferences). Most 2nd graders used a proportion strategy (d, e, f, g) with lines but a numerator strategy (a) with dots (17/27), and most 5th graders did the same (15/29), but a small group used a proportion strategy (d, e, f, g) with both (7/29). Adults showed the most variability with almost half the adult sample using a proportion strategy (d, e, f, g) for both lines and dots (14/32) and most of the remaining using a proportion strategy (d, e, f, g) when comparing lines and a numerator strategy (a) when comparing dots (13/32).

4. Experiment 3

Together, Experiments 1 and 2 provide initial evidence that the strategies people use vary as a function of the format of the stimulus used to present the proportional information. However, in both experiments, data was collected primarily for another purpose and the procedures involved additional tasks and features that are unnecessary – or potentially confounding – for our purposes. Thus, we collected new data with more trials that better balance a range of proportion comparisons, include two sets of continuous and discrete stimuli, and pre-registered our model comparison analysis.

4.1. Method

4.1.1. Participants

One hundred and fifty-seven adults ($M_{\text{age}} = 39$ years, Range: 21 to 70 years (4 missing data); 69 women, 84 men, 2 nonbinary, 2 missing data) are included in the analyses. Three additional adults participated and were excluded following our pre-registered exclusion criteria (see Data Analysis). Our analyzed sample was 68 % white and not Hispanic or Latine, 14 % Black or African American, 6 % Asian, 7 % Hispanic or Latine, and 5 % more than one race and/or ethnicity. Most participants were educated, with only 34 % having less than a Bachelor's degree (or equivalent), 41 % having a Bachelor's degree (or equivalent), and 25 % having more than a Bachelor's degree (e.g., some graduate training, a graduate degree).

Adults participated entirely online and were recruited from Prolific. The study lasted approximately 12 min and adults were compensated $2.40 (for a rate of about $12/h). Adults were randomly assigned to see either area-based proportion stimuli (pie charts and blobs, order counterbalanced; $N = 78$) or number-based proportion stimuli (dots, with and without area control, order counterbalanced; $N = 79$).

4.1.2. Stimuli and materials

There were four types of stimuli: blobs, pie charts, equal sized dots, and varied dots. Each participant completed two blocks. Participants assigned to see area-based proportion completed a block of blobs and a block of pie charts (order counterbalanced). Participants assigned to see number-based proportion completed a block of equal sized dots, without controlling for cumulative area, and varied dots that did control for cumulative area (order counterbalanced). Each block included 120 trials[2] (60 unique proportion comparisons, shown once with the correct answer on the right and once with the correct answer on the left), presented in a random order.

Identical proportion comparisons were used across all four stimulus types and each proportion was defined as the proportion of the shape (area-based) or set of dots (number-based) that was orange (or blue, target color was randomly assigned and counterbalanced across participants). Proportion comparisons were selected so that on half the trials the stimulus with the larger proportion (i.e., the correct answer) also had the larger numerator (e.g., 28/52 vs 33/43) and on the remaining trials the stimulus with the larger proportion had the smaller numerator (e.g., 13/33 vs 8/14). The ratio between the two proportions (i.e., larger proportion/smaller proportion) ranged from 1.005 to 2.467, $M = 1.44$. The ratio between the two numerators ranged from 1.026 to 3.75, $M = 1.52$.

Number-based proportion stimuli were presented as orange and blue dots intermixed on a grey background. The number of dots of a single color ranged from 4 to 41 and the total number of dots ranged from 14 to 53. In one block of trials, the dots were all equal in size (radius = 10 pixels), so that the proportion of orange (or blue) area was equal to the proportion based on number. In the other block of trials, the area of orange and blue within a stimulus was approximately equal so that area-based proportion could not be used as a cue for number-based proportion judgements. To equate area, we varied the size of the dots allowing for up to five distinct radii.

Area-based proportion stimuli were presented as a shape (blob or circle, in two separate blocks) partially colored orange and partially colored blue, on a grey background. The pie chart stimuli were presented as a circle divided into two segments (one orange, one blue). The blob stimuli were presented as an irregular blob with a straight line dividing the blob into two portions (one orange, one blue). In both cases, the absolute location of the stimulus and the orientation of the stimulus was random so that the relative position of the orange and blue portions and the absolute position of the shape within the grey rectangle were not predictable. To equate proportion comparisons across stimuli formats, each unit was given a value equal to the area of a single dot in the equal sized dot stimuli described above. For example, a fraction of 3/10 would have three orange dots and seven blue dots (or vice versa) in the number-based stimuli and an area of approximately 942 orange pixels

[2] Because of an error, one trial included the same quantity as the non-numerator component and was excluded from analyses since model b would not apply to that trial.

and 2199 blue pixels (or vice versa) in the area-based stimuli.

4.1.3. Procedure

Participants were randomly assigned to see either both area-based proportion stimuli or both number-based proportion stimuli. The procedure within each condition and each block was identical and all that differed was the format of the stimuli. Written instructions were provided on the screen prior to each block. Participants were told to compare the proportion of the set of dots, pie chart, or blob (based on block and condition) that is orange (or blue, counterbalanced). Participants responded by pressing the right or left arrow key for the right or left stimulus, respectively, and were told to respond as quickly as possible. Stimuli remained visible for 1200 ms and adults were given unlimited time to respond. Prior to completing the task, participants were given four practice trials with accuracy feedback (i.e., on-screen text telling them they were correct or incorrect). On test trials, participants were not given accuracy feedback and between trials were presented with a visual mask (1000 ms).

On 8 trials per block, immediately after responding participants were asked to rate their confidence making the judgement immediately prior. Participants rated their confidence on a continuous slider scale only marked with end points of "Very Unsure" (scored as 0) and "Very Confident" (scored as 100). The same eight proportion comparisons were selected for all participants so that comparisons with various properties and levels of difficulty were probed. Note that the randomized order of the trials means that when in the sequence participants were asked was also randomized.

After the primary task, participants were asked whether they were paying attention (multiple choice options: yes, sometimes, no), to recall what they were asked to do (free-response text box), to describe the strategy they used (free-response), and to provide any comments or concerns about the study.

The task was programed in jsPsych (de Leeuw et al., 2023) and presented to participants via Pavlovia (https://pavlovia.org/).

4.1.4. Analytical details

As pre-registered, we excluded individual trials that had reaction times less than 200 ms (139/37680 trials, i.e., 0.004 %) and entire participants that reported not paying attention to the study ($N = 0$) or who had more than 50 % of their trial-level data excluded ($N = 3$).

Data was analyzed as described for Experiments 1 and 2. Model diagnostics showed 1 divergent transition, appropriately large effective sample sizes, and all $\widehat{R} = 1$, suggesting the models did converge.

4.1.5. Transparency and openness

The sample size, design, hypotheses, and analysis plan were pre-registered (https://osf.io/4g7fq). Additional exploratory analyses involving confidence judgements are also included and are specifically described as exploratory. All stimuli, materials, data, and analysis scripts are publicly available (https://osf.io/2rtdq). Participants provided

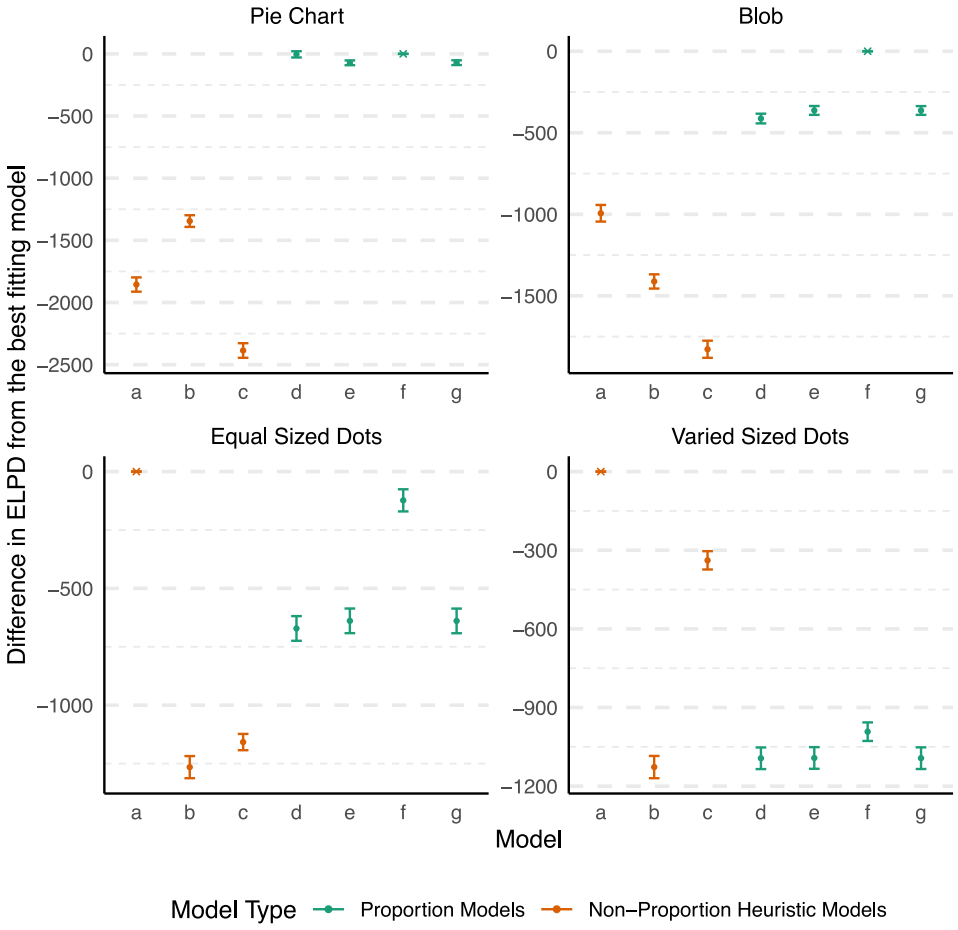

Fig. 8. Differences in Expected Log Predictive Densities (ELPD) from the best fitting model, separately for each format. The best fitting model is represented as an x and error bars are estimated standard errors of the difference from this model. Non-proportion heuristic models (a, b, c) are in orange and proportion models (d, e, f, g) are in green (see Table 1 for details of each lettered model). a = Weber-Dependent Comparison of Numerators; b = Weber-Dependent Comparison of Non-Numerator Components; c = Weber-Dependent Comparison of Denominators; d = Weber Dependent Comparison of Proportions; e = One-Cycle Power Model of Proportions; f = Beta Binomial Model of Proportions; g = Logistic Model on the Difference Between Proportions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

informed consent and the study was approved by Rutgers University IRB (Pro2023001930, "Characterizing Quantiative Development").

4.2. Results and discussion

In Experiment 3, we used the same analytical strategy, with data from four stimuli formats: two area-based continuous and two number-based discrete. Point estimates of the model parameters and model comparison metrics are presented in Supplemental Materials A and B, respectively. Differences in ELPD values are presented in Fig. 8 and comparisons of model predictions versus empirical accuracy are presented in Fig. 9.

For pie chart comparisons, the Beta Binomial proportion model (f; which numerically best fit the data) and the Weber Proportion model (d) were almost identical (D = 0.15), and the remaining proportion models were also indistinguishable (f vs. e, g: Ds < 4). In contrast, all three non-proportional heuristic models (a, b, c) were substantially worse fitting (all Ds > 28). For blob comparisons, the Beta Binomial proportion model (f) best fit the data and all other models were substantially worse fitting, both the other proportion models (f vs. d, e, g; Ds > 13) and the non-proportional heuristic models (f vs. a, b, c; Ds > 19).

When comparing sets of equal sized dots, the numerator model (a) best fit the data, and was indistinguishable from the Beta Binomial proportion model (f; D = 2.59). All other models, both proportional and

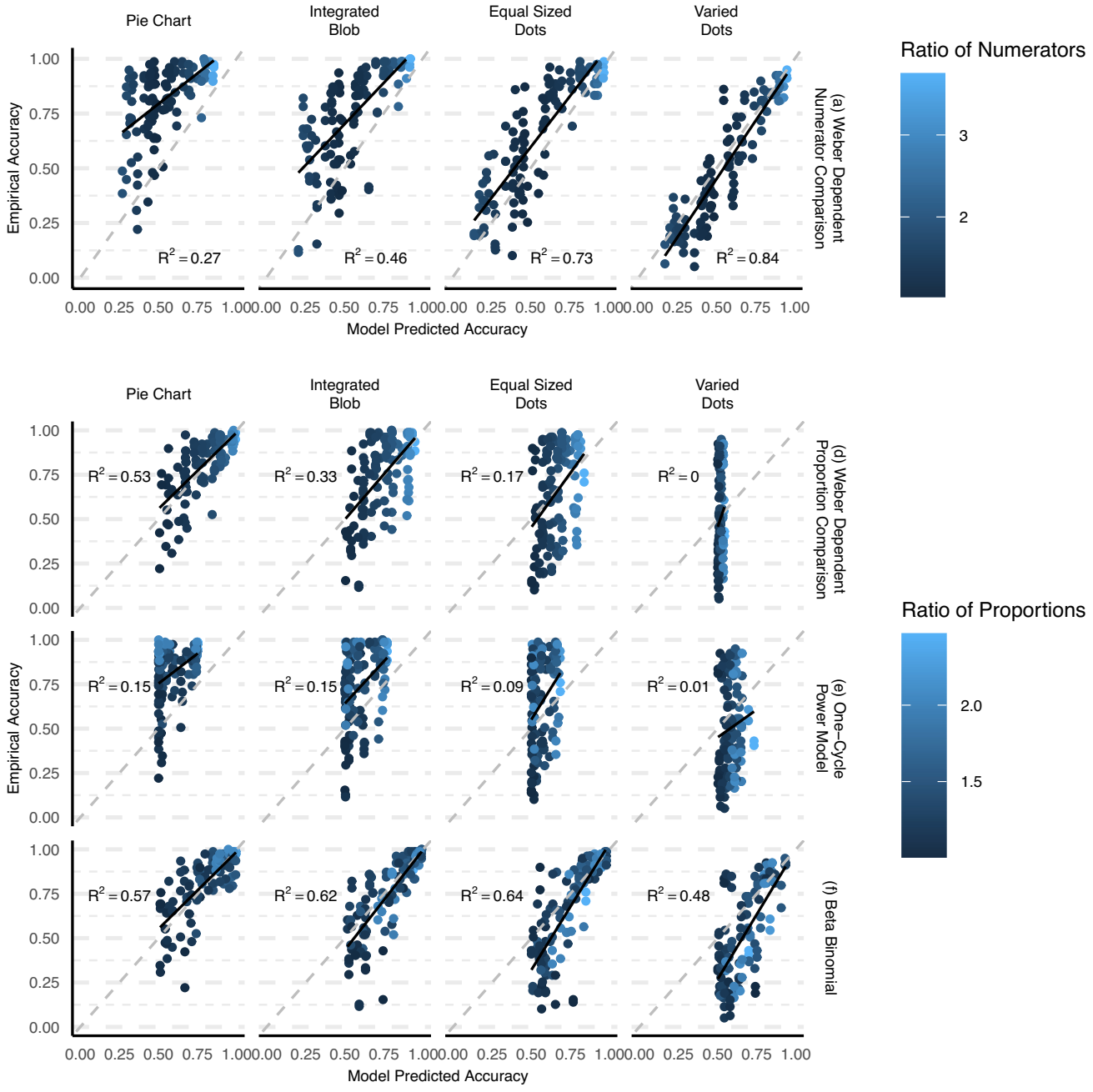

Fig. 9. Model predicted probability correct (x-axis) and empirical accuracy (proportion correct; y-axis) at the trial level from Experiment 3, colored as a function of the ratio between the numerators (larger numerator/smaller numerator) for the numerator comparison model (a) and between the proportions (larger proportion/smaller proportion) for the proportion models (d, e, f). Model estimates are based on models using the parameter point-estimates for each format separately. Each point is summary data across participants for a single trial. The dotted diagonal line is along x = y, as an indicator of what perfect agreement would look like. The R^2 estimates were calculated from linear regression models predicting empirical accuracy from model predicted accuracy, summarized at the trial level.

non-proportional heuristic models, were substantially worse fitting (all Ds > 12). When comparing sets of dots that varied in size (and equated cumulative area), the numerator model (a) best fit the data and all other models were substantially worse fitting (all Ds > 9).

Parameter estimates were similar to those reported for adults in Experiments 1 and 2 (see Supplemental Materials A). The Weber parameter estimate from the numerator comparison model (a) when comparing dots was again higher than typically reported for absolute number comparison tasks (equal sized $w = 0.38$ and varied dots $w = 0.43$). For area-based comparisons, the proportion Weber estimate (d) was in a similar range (Pie Chart: $w = 0.23$, Blob: $w = 0.25$) and the beta parameter estimates in the one-cycle power model (e) were also slightly above one (Pie Charts: beta = 1.10, Blob: 1.22).

When looking at each individual separately (see Supplemental Materials C), the same general pattern was found, but again with substantial variability and smaller differences in ELPD between models (i.e., the models were less distinguishable at the individual level). In fact, for none of the participants was there more than three standard errors between the best fitting model and the other models. For completeness and consistency with the other studies, we report the best fitting models based on the ordinal pattern. Most adults were best fit by a proportion model (d, e, f, g) when comparing pie charts (72/78), with the remaining best fit by the numerator model (a; 6/78). Similarly, when comparing blobs, most adults were best fit by a proportion model (53/78), though a sizeable number were best fit by the numerator model (a; 25/78). In contrast, when comparing equal sized dots, most adults were best fit by the numerator model (49/79), with the remaining mostly best fit by a proportion model (26/79). When comparing varied dots, most adults were best fit by the numerator model (37/79) or the denominator model (24/79), with the remaining best fit by a proportion model (13/79) or the non-numerator component model (5/79).

Finally, we explored adults' confidence ratings across different stimuli and when the numerator and proportion were congruent (i.e., the larger proportion also had the larger numerator) or incongruent (i.e., the larger proportion had the smaller numerator; see Fig. 10). Overall, adults were most confident when comparing pie charts (vs. all others, $p < .01$), followed by blobs (vs. both dot condition, $ps < 0.05$), and equal sized and varied dots, which were not significantly different ($p = .732$). Additionally, when making judgements about pie charts ($M_{congr} = 75.6$, $M_{incongr} = 71.2$), blobs ($M_{congr} = 70.2$, $M_{incongr} = 62.7$), and equal sized dots (i.e., not area controlled; $M_{congr} = 63.5$, $M_{incongr} = 58.0$), but not varied dots ($M_{congr} = 60.1$, $M_{incongr} = 59.5$), adults were more confident on numerator congruent trials than numerator incongruent trials ($ps < 0.01$; varied dots, $p = .627$).

5. General discussion

We used a Bayesian model comparison approach to investigate differences in strategy use when people were asked to compare proportions in different formats. Using three datasets generated from tasks with different instructions and stimuli, but the same general paradigm structure, our results (summarized in Table 2) reveal systematic differences in strategy use across discrete versus continuous displays of proportion, with additional nuances *within* discrete and continuous stimuli. In general, we found that people tended to use a proportion comparison strategy when asked to compare continuous proportion but were more likely to use a numerator comparison strategy when asked to compare discrete proportion. This suggests that people are drawing upon fundamentally different strategies that rely on different information when asked to compare proportions in discrete and continuous formats, even though the task instructions, structure, and goal remained constant (within an experiment). To be clear, we do not interpret these findings as suggesting that future work on proportional reasoning should exclusively use part-whole, area-based stimuli as a "true" measure of proportion. Instead, we suggest that domains involving proportional reasoning should systematically incorporate variation (e.g., different stimuli, different contexts) into their paradigms and use strategy discovery methods to move beyond overall behavior and toward thinking about variation in the underlying strategies and processes people use. In doing so here, we provide a potential framework to understand the contradictory and surprising results in the proportional reasoning literature.

One motivating puzzle for this approach is that infants seem to have proportional reasoning abilities that they can use to make inferences about the world around them (e.g., Denison and Xu, 2012; McCrink and Wynn, 2007), yet older children and adults demonstrate substantial difficulty with proportional information (e.g., Boyer et al., 2008; Hurst et al., 2021). One possible explanation is that across development,

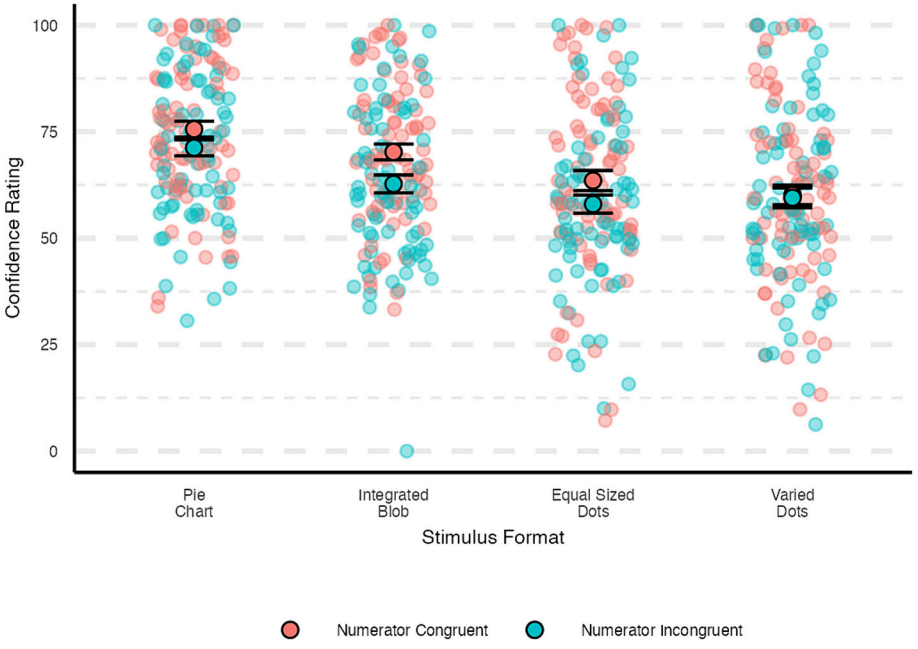

Fig. 10. Mean confidence rating on each stimulus format across numerator congruent trials (e.g., 32/39 vs. 27/50) and numerator incongruent trials (e.g., 11/20 vs. 17/47).

Table 2
Summary of results across all three experiments, grouped by stimulus category.

		Proportion Strategy	Numerator Strategy	Inconclusive and/or a mixture of strategies
Continuous Stimuli	Pie Charts	E1: $R^2 = 0.38$ E3: $R^2 = 0.57$		
	Integrated Blobs	E3: $R^2 = 0.62$		
	Separated Line Ratios	E2, 2nd: $R^2 = 0.31$ E2, 5th: $R^2 = 0.40$ E2, adults: $R^2 = 0.64$		E2, PreK: $R^2_{num} = 0.44$ $R^2_{best\ prop} = 0.53$
	Separated Circle Ratios	E2, adults: $R^2 = 0.59$		E2, PreK: $R^2_{num} = 0.69$; $R^2_{best\ prop} = 0.46$ E2, 2nd: $R^2_{num} = 0.70$; $R^2_{best\ prop} = 0.24$ E2, 5th: $R^2_{num} = 0.60$; $R^2_{best\ prop} = 0.48$
	Separated Blob Ratios	E2, 2nd: $R^2 = 0.59$ E2, 5th: $R^2 = 0.54$ E2, adults: $R^2 = 0.51$		E2, PreK: $R^2_{num} = 0.47$; $R^2_{best\ prop} = 0.27$
	Integrated, Equal Sized Dots			E3: $R^2_{num} = 0.73$; $R^2_{best\ prop} = 0.64$
	Integrated Varied Dots		E1: $R^2 = 0.76$ E3: $R^2 = 0.84$	
Discrete Stimuli				E2, PreK: $R^2_{num} = 0.58$; $R^2_{best\ prop} = 0.16$
	Separated Dot Ratios		E2: 2nd: $R^2 = 0.73$	E2, 5th: $R^2_{num} = 0.69$; $R^2_{best\ prop} = 0.30$ E2, adults: $R^2_{num} = 0.70$; $R^2_{best\ prop} = 0.30$

Note: E1 = Experiment 1, E2 = Experiment 2, E3 = Experiment 3.
For E2, age groups are separated as PreK = preschool, 2nd = 2nd graders, 5th = 5th graders, and adults.
For each dataset, the adjusted R^2 from the linear regression predicting empirical accuracy from model predicted accuracy on each trial is reported for the best fitting model, as determined by our model comparison approach using $ELPD_{diff}$. When the $ELPD_{diff}$ between the numerator model and the best fitting proportion model was within five standard errors, the results are categorized as inconclusive and/or a mixture of strategies, and the R^2 from both the estimated numerator model and best fitting proportion model are reported.

people use different strategies – even within the same stimulus format. In the current study, we found very similar patterns across 4- to 5-year-olds, 2nd graders, 5th graders, and adults. The general pattern of using proportion strategies for continuous stimuli (in particular, line lengths) and using numerator strategies, or a mix of numerator and proportion strategies, for discrete stimuli was found across age groups. These results highlight that some of children's and adults' behavioral difficulties with discrete proportion tasks may be explained by them using non-proportional heuristic strategies, rather than due to poor proportional reasoning, per se. With our current data, we cannot speak to whether differences in strategy use across development may also explain some of the puzzling developmental patterns, though this is an important question for future work. In other words, it may be that rather than people experiencing a decrease in proportional reasoning ability with age (something that would be surprising), people are learning different kinds of strategies and when to use them, resulting in them misapplying a heuristic in commonly used discrete contexts (Boyer et al., 2008). Importantly, however, it may be that this incorrect heuristic-based strategy is not used early in development. More work is needed to test this account of the developmental data, including applying a strategy-based analysis to data from a much wider age range, both before and after we expect children to have learned certain strategies.

Although we primarily focused on the distinction between discrete proportion stimuli and continuous proportion stimuli, it is important to note that discreteness, per se (i.e., the availability of numerical information) is not the only difference between the discrete and continuous stimuli used here. The discrete stimuli in Experiments 1 and 3 were composed of intermixed dots of varying sizes (to de-confound cumulative area). It may be that the intermixing of the two components or the irregularity of the different sized dots impact strategy use, rather than or in addition to the discreteness. In fact, when the dots were equated in size, allowing area and number to be confounded in Experiment 3, we see a mix of numerator and proportion strategies, Furthermore, there were also differences in strategy use within the "continuous" stimuli. In Experiment 2, younger children showed a different pattern of strategy use for circles and blobs compared to lines, suggesting that there may be some feature of the circle and blob comparisons (e.g., the vertical arrangement) that evoked numerator strategies, in addition to proportion strategies. Additionally, in Experiment 3, adults' data were best fit by different formalizations of the proportion strategies for blobs versus pie charts, which may suggest a different psychological process underlying the proportion comparison strategies. Together, these results lead to a more general argument that people's strategies differ as a function of the stimulus and that these different strategies may be evoked by conceptual distinctions (e.g., the availability, or not, of numerical information) and perceptual distinctions (e.g., spatial organization and structure of the information). Exploratory analyses involving adults' subjective confidence also raise the question of whether adults might be modulating their strategies because of differences in perceived difficulty. Although we cannot speak to a causal pattern with the current data, it may be that adults switch to a heuristic based numerator strategy in cases where they are less confident in their judgements, as was the case for the discrete dot comparisons in Experiment 3.

Beyond model comparison, the parameter estimates in each of the best fitting models provide insight into interpreting these strategies. The Weber estimates from the numerator comparison strategy on discrete data decreased across development, as we would expect, but were higher than those typically reported on absolute number comparison tasks (e.g., Odic et al., 2013). This might suggest that comparing absolute numbers in this context was more difficult than in other tasks designed specifically to measure the approximate number system. Alternatively, it may be that the numerator comparison model alone is

not a perfect model of behavior on this task. For example, it may be that people's comparisons are highly influenced by the numerator value but in a way that also incorporates other quantitative information about the display, resulting in a hybrid model (Alonso-Díaz et al., 2018; Braithwaite and Siegler, 2018). Another possibility is that there is additional variability across people or trials that is not well captured here. Our analysis of individuals provides support for this account, revealing that some people were similarly fit by both the proportion models and the numerator heuristic model, which may be due to using different strategies on different trials.

For continuous stimuli, the parameters on each proportion model were generally consistent across the experiments for adults and showed the developmental patterns we would expect. It is worth noting that although the beta parameter estimates from the one-cycle power models were below one for children, they were above one for adults, which is surprising given prior work revealing estimates below one for categorically similar stimuli in adults (Hollands and Dyre, 2000). However, these prior psychophysical estimates were generated from estimation tasks, not comparison tasks. Here, we assume people use the one-cycle power model as an estimation strategy, but we used a logistic model to approximate the comparison process. It is possible that this parameter estimate is biased because it is incorporating additional noise in the comparison process that is not well captured by our logistic model. Together, the similar model fits across proportion strategies and the interpretation of these parameter estimates highlight the need for substantially more psychophysical and computational research on proportional estimation and comparison across different quantitative displays.

It is important to note three (non-exhaustive) limitations of our modeling approach. First, we could not differentiate between most of the proportion models. One possibility is that the comparison context, as opposed to estimation contexts that have been historically used in the psychophysical literature of numerical and proportional reasoning, is less sensitive to the small differences in each of the proportion models. To better differentiate these strategies, future work could create a more specific set of stimuli that capitalizes on the cases where the differences in noise and bias in each model would make different predictions, as in approaches to optimal experimental design (Smucker et al., 2018). Second, our results focused primarily on which model fit the data better, but we did not provide an estimate of absolute model fit. Furthermore, visualizations comparing empirical data and model predictions provide some insight into this issue and reveal that for some formats and age groups, it seems like none of the models provided a particularly compelling (qualitative) fit with the data. It may be that alternative proportion models or a hybrid model that incorporates both absolute components and proportion magnitude are necessary for better explaining people's behavior across formats. Finally, despite increasing the number of trials in Experiment 3, we were unable to provide high quality model fits that disambiguated different strategies at the individual level. It is likely that there are substantial individual differences in strategy use, and that at least some participants use multiple strategies within a session. More individual data, and potentially different analytical methods, are needed to better capture this individual variability.

Thus, systematically manipulating the perceptual and conceptual structure of the proportion stimuli, beyond just discreteness, while also capturing individual strategy use and the relations with more general individual and developmental differences (e.g., math ability, executive functioning, spontaneous focusing on number), is necessary to get a better picture of why some formats evoke numerator heuristic strategies in some people and at some points in development.

When proportional information is presented with different types of quantities people use different strategies, even to complete the same task of comparing proportions. These findings suggest that the typical goal of investigating *the* cognitive strategy underlying people's approach to proportion tasks is likely misguided because the fundamental assumption - that there is a single cognitive strategy to find - is incorrect. Instead, behavioral differences may stem from the use of different strategies entirely. This explanation requires pairing computational strategy discovery methods with human experiments to systematically investigate the factors that prompt the use of some strategies over others. Our approach joins a recent and growing call for considering heterogeneity in cognitive science (Bryan et al., 2021; Lewis, 2022; Newcombe et al., 2022), and provides a framework for considering heterogeneity in strategy use across problem contexts, even within the same individuals.

CRediT authorship contribution statement

Michelle A. Hurst: Writing – original draft, Visualization, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation. **Steven T. Piantadosi:** Writing – review & editing, Software, Resources, Methodology, Conceptualization.

Declaration of competing interest

None.

Data availability

All data and code, and most materials are available here: https://osf.io/2rtdq. Materials for Experiments 1 and 2 and pre-registrations for Experiments 1 and 3 are linked separately (links in manuscript).

Acknowledgements

We would like to thank Susan C. Levine for financial and institutional support in collecting data for Experiment 1 and Yunji Park for providing information about data used in Experiment 2. Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number K99HD104990 and R00HD104990 to MAH. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2024.105918.

References

[dataset] Author. (). *Strategies Across Proportion Displays. Open Science Framework*. https://osf.io/2rtdq. last updated 2024-08-05. https://doi.org/10.17605/OSF.IO/2RTDQ.

Alonso-Díaz, S., Piantadosi, S. T., Hayden, B. Y., & Cantlon, J. F. (2018). Intrinsic whole number bias in humans. *Journal of Experimental Psychology: Human Perception and Performance, 44*(9), 1472–1481. https://doi.org/10.1037/xhp0000544

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods, 52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Boyer, T. W., Levine, S. C., & Huttenlocher, J. (2008). Development of proportional reasoning: Where young children go wrong. *Developmental Psychology, 44*(5), 1478–1490. https://doi.org/10.1037/a0013110

Braithwaite, D. W., & Siegler, R. S. (2018). Developmental changes in the whole number bias. *Developmental Science, 21*(2), Article e12541. https://doi.org/10.1111/desc.12541

Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour, 5*(8), 980–989. https://doi.org/10.1038/s41562-021-01143-3

Bryant, P., & Nunes, T. (2012). *Children's understanding of probability: A literature review (summary report)*. Nuffield Foundation.

Cook, J. D. (2012). Fast approximation of Beta inequalities. http://biostats.bepress.com/mdandersonbiostat/paper76.

Denison, S., & Xu, F. (2010). Twelve- to 14-month-old infants can predict single-event probability with large set sizes: Twelve- to 14-month-olds can predict single-event probability. *Developmental Science, 13*(5), 798–803. https://doi.org/10.1111/j.1467-7687.2009.00943.x

Denison, S., & Xu, F. (2012). Probabilistic inference in human infants. In , *Vol. 43*. *Advances in child development and behavior* (pp. 27–58). Elsevier. https://doi.org/10.1016/B978-0-12-397919-3.00002-2.

Denison, S., Reed, C., & Xu, F. (2013). The emergence of probabilistic reasoning in very young infants: Evidence from 4.5- and 6-month-olds. *Developmental Psychology, 49* (2), 243–249. https://doi.org/10.1037/a0028278

DeWolf, M., Bassok, M., & Holyoak, K. J. (2015). Conceptual structure and the procedural affordances of rational numbers: Relational reasoning with fractions and decimals. *Journal of Experimental Psychology: General, 144*(1), 127–150. https://doi.org/10.1037/xge0000034

Fabbri, S., Caviola, S., Tang, J., Zorzi, M., & Butterworth, B. (2012). The role of numerosity in processing nonsymbolic proportions. *The Quarterly Journal of Experimental Psychology, 65*(12), 2435–2446. https://doi.org/10.1080/17470218.2012.694896

Faulkenberry, T. J., & Pierce, B. H. (2011). Mental representations in fraction comparison: Holistic versus component-based strategies. *Experimental Psychology, 58* (6), 480–489. https://doi.org/10.1027/1618-3169/a000116

Fazio, L. K., DeWolf, M., & Siegler, R. S. (2016). Strategy use and strategy choice in fraction magnitude comparison. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(1), 1.

Gallistel, C. R., & Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition, 44*(1–2), 43–74.

Gillard, E., Van Dooren, W., Schaeken, W., & Verschaffel, L. (2009). Proportional reasoning as a heuristic-based process: Time constraint and dual task considerations. *Experimental Psychology, 56*(2), 92–99. https://doi.org/10.1027/1618-3169.56.2.92

Girotto, V., Fontanari, L., Gonzalez, M., Vallortigara, G., & Blaye, A. (2016). Young children do not succeed in choice tasks that imply evaluating chances. *Cognition, 152*, 32–39. https://doi.org/10.1016/j.cognition.2016.03.010

Hoffman, M. D., & Gelman, A. (2014). The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research, 15*, 31.

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

Hurst, M. A., & Cordes, S. (2016). Rational-number comparison across notation: Fractions, decimals, and whole numbers. *Journal of Experimental Psychology: Human Perception and Performance, 42*(2), 281–293. https://doi.org/10.1037/xhp0000140

Hurst, M. A., & Cordes, S. (2018). Attending to relations: Proportional reasoning in 3- to 6-year-old children. *Developmental Psychology, 54*(3), 428–439. https://doi.org/10.1037/dev0000440

Hurst, M. A., & Levine, S. C. (2022). Children's understanding of most is dependent on context. *Cognition, 225*, Article 105149. https://doi.org/10.1016/j.cognition.2022.105149

Hurst, M. A., & Piantadosi, S. (2022). Investigating Adults' Strategy Use During Proportional Comparison. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (p. 44).

Hurst, M. A., Shaw, A., Chernyak, N., & Levine, S. C. (2020). Giving a larger amount or a larger proportion: Stimulus format impacts children's social evaluations. *Developmental Psychology, 56*(12), 2212–2222. https://doi.org/10.1037/dev0001121

Hurst, M. A., Boyer, T. W., & Cordes, S. (2021). *Spontaneous and directed attention to number and proportion*. Journal of Experimental Psychology: Learning, Memory, and Cognition. https://doi.org/10.1037/xlm0001084

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. https://doi.org/10.1080/15248370701202471

Kalra, P. B., Binzak, J. V., Matthews, P. G., & Hubbard, E. M. (2020). Symbolic fractions elicit an analog magnitude representation in school-age children. *Journal of Experimental Child Psychology, 195*, Article 104844. https://doi.org/10.1016/j.jecp.2020.104844

Lamon, S. J. (1993). Ratio and proportion: Connecting content and Children's thinking. *Journal for Research in Mathematics Education, 24*(1), 41. https://doi.org/10.2307/749385

Lewis, N. A. (2022). What would make cognitive science more useful? *Trends in Cognitive Sciences, S1364661322001620*. https://doi.org/10.1016/j.tics.2022.07.005

McCrink, K., & Wynn, K. (2007). Ratio abstraction by 6-month-old infants. *Psychological Science, 18*(8), 740–745. https://doi.org/10.1111/j.1467-9280.2007.01969.x

Meck, W. H., & Church, R. M. (1983). A mode control model of counting and timing processes. *Journal of Experimental Psychology: Animal Behavior Processes, 9*(3), 320.

Newcombe, N. S., Hegarty, M., & Uttal, D. (2022). Building a cognitive science of human variation: Individual differences in spatial navigation. *Topics in Cognitive Science*. https://doi.org/10.1111/tops.12626

Obersteiner, A., Alibali, M. W., & Marupudi, V. (2022). Comparing fraction magnitudes: Adults' verbal reports reveal strategy flexibility and adaptivity, but also bias. *Journal of Numerical Cognition, 8*(3), 398–412. https://doi.org/10.5964/jnc.7577

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

Park, Y., Viegut, A. A., & Matthews, P. G. (2020). More than the sum of its parts: Exploring the development of ratio magnitude versus simple magnitude perception. *Developmental Science*. https://doi.org/10.1111/desc.13043

Piaget, J., & Inhelder, B. (1975). *The origins of the idea of chance in children*. Norton.

Piantadosi, S. T. (2016). Efficient estimation of Weber's W. *Behavior Research Methods, 48* (1), 42–52. https://doi.org/10.3758/s13428-014-0558-8

Placì, S., Fischer, J., & Rakoczy, H. (2020). Do infants and preschoolers quantify probabilities based on proportions? *Royal Society Open Science, 7*, Article 191751. https://doi.org/10.1098/rsos.191751

R Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing https://www.R-project.org/.

Schneider, M., & Siegler, R. S. (2010). Representations of the magnitudes of fractions. *Journal of Experimental Psychology. Human Perception and Performance, 36*(5), 1227.

Shuford, E. H. (1961). Percentage estimation of proportion as a function of element type, exposure time, and task. *Journal of Experimental Psychology, 61*(5), 430–436. https://doi.org/10.1037/h0043335

Siegler, R. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology: General, 116*, 250–264. https://doi.org/10.1037/0096-3445.116.3.250

Siegler, R. S. (1991). Strategy choice and strategy discovery. *Learning and Instruction, 1* (1), 89–102. https://doi.org/10.1016/0959-4752(91)90020-9

Siegler, R. S. (1994). Cognitive variability: A key to understanding cognitive development. *Current Directions in Psychological Science, 3*(1), 1–5.

Smucker, B., Krzywinski, M., & Altman, N. (2018). Optimal experimental design. *Nature Methods, 15*(8). https://doi.org/10.1038/s41592-018-0083-2. article 8.

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

Stan Development Team. (2020). RStan: The R interface to Stan (R pacakge version 2.21.2). http://mc-stan.org/.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review, 64*(3), 153.

Studio Team, R. (2016). *RStudio: Integrated development for R*. RStudio Inc.. http://www.rstudio.com/

Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *Science, 332* (6033), 1054–1059.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131.

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., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., & Gelman, A. (2022). Loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. https://mc-stan.org/loo/.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag.

Wickham, H. (2017). Tidyverse: Easily install and load the "Tidyverse". R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse.

Xu, F. (2019). Towards a rational constructivist theory of cognitive development. *Psychological Review, 126*(6), 841–864. https://doi.org/10.1037/rev0000153

Xu, F., & Denison, S. (2009). Statistical inference and sensitivity to sampling in 11-month-old infants. *Cognition, 112*(1), 97–104. https://doi.org/10.1016/j.cognition.2009.04.006

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences, 105*(13), 5012–5015. https://doi.org/10.1073/pnas.0704450105