Cognitive Psychology 122 (2020) 101309

Contents lists available at ScienceDirect

Cognitive Psychology

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

Using Occam's razor and Bayesian modelling to compare discrete and continuous representations in numerosity judgements

Jake Spicer[a,*], Adam N. Sanborn[a], Ulrik R. Beierholm[b]

[a] University of Warwick, UK
[b] Durham University, UK

ARTICLE INFO

Keywords:
Numerosity
Estimation
Rational modelling

ABSTRACT

Previous research has established that numeric estimates are based not just on perceptual data but also past experience, and so may be influenced by the form of this stored information. It remains unclear, however, how such experience is represented: numerical data can be processed by either a continuous analogue number system or a discrete symbolic number system, with each predicting different generalisation effects. The present paper therefore contrasts discrete and continuous prior formats within the domain of numerical estimation using both direct comparisons of computational models of this process using these representations, as well as empirical contrasts exploiting different predicted reactions of these formats to uncertainty via Occam's razor. Both computational and empirical results indicate that numeric estimates commonly rely on a continuous prior format, mirroring the analogue approximate number system, or 'number sense'. This implies a general preference for the use of continuous numerical representations even where both stimuli and responses are discrete, with learners seemingly relying on innate number systems rather than the symbolic forms acquired in later life. There is however remaining uncertainty in these results regarding individual differences in the use of these systems, which we address in recommendations for future work.

1. Introduction

In many everyday tasks, we are required to make quick estimates of discrete stimuli based on noisy perceptual data: the number of people in a crowded room, or cars in a lane of traffic, for example. These decisions are not solely reliant on perceptual information, but also use past experiences with such stimuli to guide responses: if estimating the number of people in a room, the actor may consider similar occasions where that number was later provided and use this information to inform their decision. Such guidance in fact becomes increasingly valuable at higher values as people's ability to discriminate between figures decreases (Krueger, 1984; Izard & Dehaene, 2008). Accurate estimates are therefore reliant on the learning of the distribution of such figures, building representations that reflect the prevalence of these values in the real world.

The influence of such previous experience is in turn however dependent on its representation, reflecting the different forms in which numerical information could be stored. Existing research has offered two potential forms for such information in two contrasting number systems, each suggesting distinct impacts on new decisions: the approximate number system and the symbolic number system. The approximate number system refers to the innate understanding of numerosity displayed by both humans and animals in which numbers are conceptualised in a continuous analogue form (Dehaene, 2011). Storing prior experiences in this

* Corresponding author at: Department of Psychology, University of Warwick, Coventry CV4 7AL, UK.
E-mail address: jake.spicer@warwick.ac.uk (J. Spicer).

https://doi.org/10.1016/j.cogpsych.2020.101309
Received 15 March 2019; Received in revised form 21 May 2020; Accepted 23 May 2020
Available online 03 July 2020

format should therefore lead future estimates to focus on values similar to those previously seen; if the previous room contained 50 people, then nearby figures such as 49 or 51 would also become more likely (e.g. Gershman & Niv, 2013). In contrast, the symbolic number system is the discrete verbal format learned in later life which allows for more complex mathematical operations (Izard & Dehaene, 2008); in this case, only the experienced value would increase in expectancy, making that response alone more likely in subsequent estimates. Such a representation would allow the learner to acquire reasonably complex distributions through experience, tracking the individual appearance rate of each potential value (e.g. Sanborn & Beierholm, 2016). This would, however, also be possible using a sufficiently complex continuous format: narrow similarity functions could emulate discrete formats, making it difficult to distinguish between these forms.

This then raises the question of which of these systems underlies discrete estimates: symbolic representations could be used to suit the discrete nature of responses and feedback, while continuous forms may be used in spite of these elements to suit the more analogue perceptual data and translated into discrete figures as required. Despite the impact of this distinction on both the representation formed and the resulting behaviour, this has received little attention in previous research. What is more, what work has been done has found conflicting results, with studies finding evidence for both continuous (Gershman & Niv, 2013) and discrete (Sanborn & Beierholm, 2016) underlying systems.

The current study therefore attempts to separate these forms using two complementary methodologies: first, an empirical contrast taking advantage of a difference in the definition of simplicity within continuous and discrete representations, and second, a quantitative contrast between computational models of behaviour in this task. In the following sections, we introduce potential models of estimation following such discrete and continuous formats, examine the principles of these models to derive methods of distinction, and use both empirical and computational comparisons to provide insight into the representations used in numeric estimates.

1.1. Using prior experience

We begin by examining the process by which past estimates could be used to inform new judgements. While this has not been studied extensively in estimation, one existing theory which touches on this process is calibration; in this theory, past trials are suggested to be used as anchoring points to map a discrete response scale onto continuous numerical representations to make subsequent estimates more accurate (Krueger, 1984; Izard & Dehaene, 2008). In this case, numerical data is automatically encoded in a continuous format and translated into discrete figures as required; for example, Izard and Dehaene (2008) suggest an affine transformation between continuous and discrete formats, using parameters to adjust both the shape and position of the discrete response scale. Calibration therefore increases the accuracy of this translation by tuning these parameters to suit the observed data, better mapping these two scales against one another to improve all future estimates. Such a transformation is, however, limited in the probability distributions it is able to represent; while this may be sufficient for reasonably simple structures, more complex distributions such as those with multiple modes cannot be accurately represented by this process alone. This stands in contrast to empirical data showing that learners can in fact acquire such multimodal distributions (Sanborn & Beierholm, 2016; Gershman & Niv, 2013). What is more, these studies also provide evidence against the use of a more complex translation function (Sanborn & Beierholm, 2016), thereby suggesting the learning of these forms is reliant on other mechanisms besides calibration. More flexible systems are therefore required to accurately represent these more complex forms, with calibration possibly offering a supporting process.

An alternative framework for the use of past experience is provided by Bayesian Decision Theory (BDT), in which prior assumptions regarding the distribution of the target stimuli are combined with direct observational data to form a posterior distribution from which a response can be selected; feedback from this response can then be used to update the representation for use in subsequent estimates. Previous observations are therefore used to inform new responses by constructing a mental representation of the true distribution, noting the prevalence of particular values. This provides BDT with an advantage over calibration as it can capture more complex learning structures such as the multimodal distributions noted above, with estimates reflecting both current perceptual data as well as the history of past observations. BDT may then provide a clear and established method well suited to the modelling of numerical estimation, better capturing the underlying process. In fact, BDT has been previously used as a description of the estimation process within continuous motor responses (Kording & Wolpert, 2004; Acerbi, Vijayakumar, & Wolpert, 2014; Chalk, Seitz, & Series, 2010), further supporting its use in the present study.

The use of BDT also facilitates the current comparison between discrete and continuous representations: while the general principles of BDT may remain fixed, the definitions of individual elements can vary, allowing for contrasts between alternate Bayesian models with different representational formats. Here, this applies primarily to the structure of the prior distribution, as this provides the assumed model of the environment, and so the representation of numerical information. The current study therefore focuses on contrasts between differing definitions of the prior, while other model elements remain identical. What is more, BDT also allows for such a distinction without specifying the algorithm used by real learners, instead only providing useful computational descriptions of behaviour (Tauber, Navarro, Perfors, & Steyvers, 2017). This places the focus on the contrast between the uses of discrete and continuous numerical formats rather than the optimality of behaviour; while such descriptions would be optimal if these priors were accurate, without certainty of the specific priors used by actual learners, or indeed the suitability of these priors to the environment, we avoid making such claims in this paper.

Continuous prior formats are provided by a number of systems, though the present study focuses on mixtures of Gaussian components due to the flexibility of such a representation, allowing for emulation of other continuous distributions. In a Gaussian mixture, observations are grouped together based on similarity to form a set of subgroups, each described by a Gaussian distribution,

which can then be combined into a single prior (Rosseel, 2002; Vanpaemel & Storms, 2008; Anderson, 1991); these have been previously used within Bayesian models of continuous estimation (e.g. Acerbi et al., 2014), providing some basis for their use as a continuous candidate in the present contrast. Such a prior holds the advantage of flexibility, being able to adjust the number of components used in the representation to best suit observed data patterns rather than using a predefined component structure; the mixture therefore offers a richer representation than parametric prior formats, able to capture more complex structures that would not be possible with a single Gaussian distribution. This flexibility has led to the application of Gaussian mixture priors to discrete estimates in spite of their continuous format; one demonstration of this is provided by Gershman and Niv (2013), in which a Gaussian mixture prior was used to model the merging of distinct categories of discrete stimuli where these categories shared similar statistical features. In this case, the Gaussian mixture is suggested to allow for simplifications of the final representation due to a prior preference for fewer components in the distribution; this could then indicate that discrete estimates may benefit from the use of a Gaussian mixture prior in terms of cognitive economy or greater generalisability. Both continuous and discrete estimates could then make use of a common underlying estimation system which is able to adapt to the needs of the task to provide the most valuable representation, considering both the accuracy and simplicity of the resulting form.

Discrete prior formats, conversely, are provided by distributions such as the categorical prior, which can be used to record the appearance rate of each observed value, relying more on memory for past observations than an inferred statistical distribution. As with the Gaussian mixture above, such a prior offers substantial flexibility, being able to discover structure in the data instead of relying on a predefined component format. Such a prior may though be better suited to numeric estimates given its greater correspondence to the discrete nature of stimuli and responses; learners could then use this prior under the assumption that this structure is more appropriate to the nature of the task. This would, however, potentially lead to differences in behaviour according to the differing world models implicitly assumed by these prior structures. To illustrate, consider the above application of simplicity according to component count from Gershman and Niv (2013) to both the Gaussian mixture and categorical priors: in the case of the Gaussian mixture prior, a preference for fewer components is assumed to lead to the merging of subgroups where possible, leading to a smaller number of broader, more varied components. Categorical components, conversely, are discrete tallies of identical value observations and cannot be broadened in this way, meaning a reduction in the number of components would instead reduce the number of values considered in the distribution. The same fundamental principle therefore leads to widely different outcomes for these two structures, with the Gaussian mixture prior considering more values in its final posterior and the categorical prior considering fewer, demonstrating the impact of this representational format on actual estimations. To return to the previous example of counting people in a room, simplicity in the discrete case means restricting responses to a limited set of answers (e.g. low/medium/high or nearest 10), while in the continuous case, responses could focus on a single mean value, but with what could be substantial departures.

It is therefore necessary to examine the prior structures used in numerical estimation to determine whether this process relies on specialised discrete formats suiting the discrete nature of this task or more general continuous forms that can be shared with other stimuli. This has in fact been previously investigated in a study by Sanborn and Beierholm (2016) in which participants performed a dot numeration task using an underlying bimodal distribution (illustrated in Fig. 1a). In this task, participants were asked to estimate the number of dots appearing on-screen, with responses being followed by direct feedback noting the true dot count, making both participant responses and task feedback discrete and definitive, so providing clear evidence of a discrete task structure. Behaviour in the experiment was then compared with Bayesian models of estimation using differing definitions of individual model elements, including a contrast between continuous and discrete prior formats using a categorical prior and a kernel density estimate. Results from this study found behaviour was better described by the categorical prior than the kernel density estimate, suggesting that participants were using a discrete prior structure in line with the discrete nature of the task.

The findings of Sanborn and Beierholm (2016) therefore indicate that discrete estimation makes use of similarly discrete elements in order to assist in constructing more precise mental representations. There is one caveat to this finding, however: while model comparisons did suggest participant behaviour was most likely to be based on the use of discrete structures, this result could also be produced by a mixture of continuous components under certain circumstances. This is due to the previously noted flexibility of the Gaussian mixture prior: by grouping similar values together, the Gaussian mixture is able to adjust the variance of its components to suit the observed data, allowing for both broad, highly varied clusters and narrow, focussed clusters. Such narrow clusters could then essentially emulate the components of a categorical prior in which all members are identical, making the component variance zero. This concern is in fact raised in the third experiment of Sanborn and Beierholm (2016), noting that such a complex Gaussian mixture could capture the true categorical structures: a mixture prior using narrow components at the modes of the distribution and a broader component across the midrange offers a reasonable approximation of the true bimodal form (illustrated in Fig. 1b). While that experiment did attempt to control for this possibility by using a quadrimodal distribution where such emulation is less precise, this only excluded a narrow set of mixture forms, while more complex structures are still possible. As such, the results of Sanborn and Beierholm (2016) can be explained in two different ways, with different implications: participants may have been using a more precise discrete prior in accordance with the discrete nature of the task, or a more flexible Gaussian mixture prior in line with that used for continuous estimates.

It is therefore necessary to distinguish between these explanations in order to determine whether discrete estimations do indeed rely on discrete structures, or whether this was simply emulated by an otherwise continuous representation. As such, the present study aimed to perform a comparison between Bayesian estimation models using either a categorical or Gaussian mixture prior in a comparable estimation task; this builds on the results of Sanborn and Beierholm (2016) by examining a full continuous mixture model rather than one possible form of this prior for a more complete contrast of these formats.

While such a comparison provides a quantitative indication as to the underlying processes of numerical estimation, we also sought

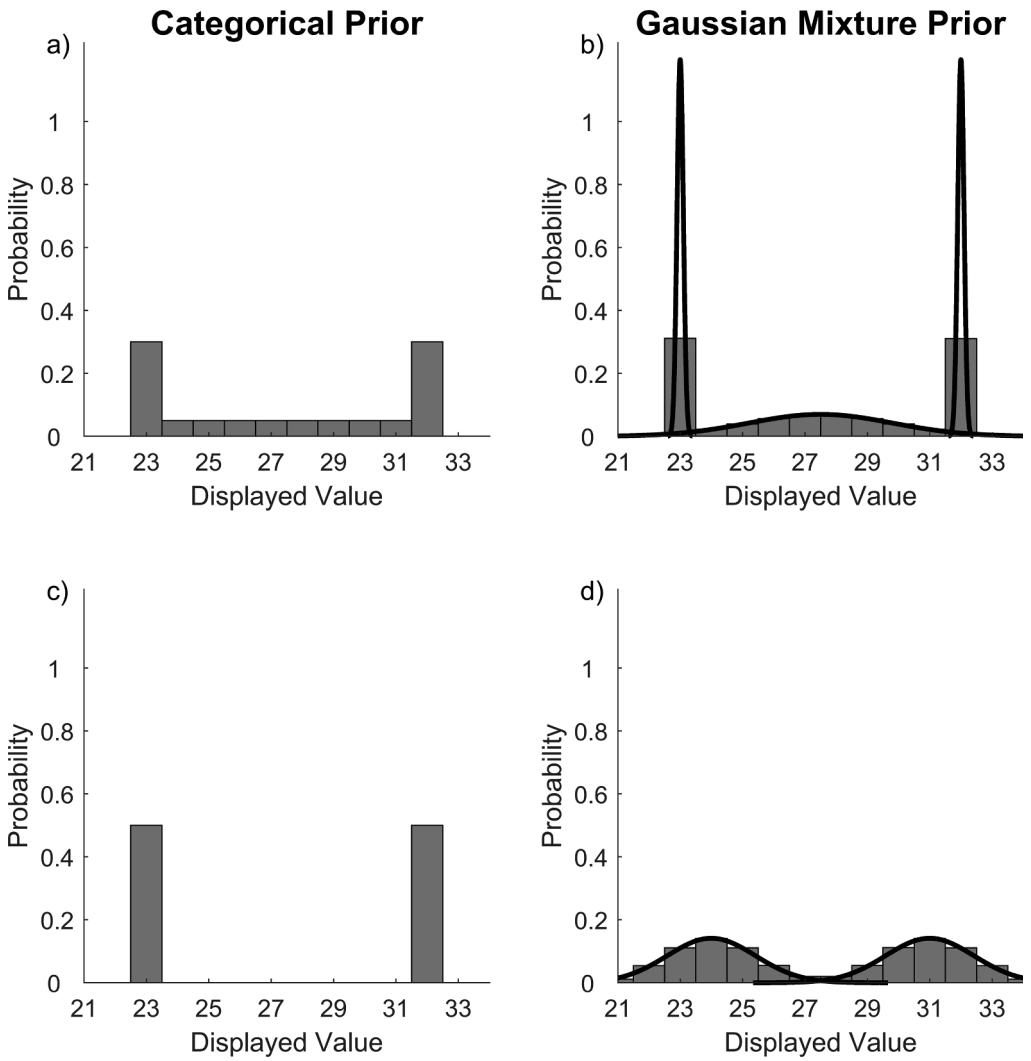

Fig. 1. Comparison of the categorical (a) and Gaussian mixture (b) priors applied to the bimodal distribution of Sanborn and Beierholm (2016). Here, the categorical matches the true distribution, and the Gaussian mixture provides an approximation, with the black lines reflecting the individual distributions of each cluster. The lower figures demonstrate the proposed simplifications of these representations via reduced component count, leading to fewer potential response values in the categorical (c), but greater bleed-over in the Gaussian mixture (d).

to supplement this contrast with a more qualitative investigation; this was intended to provide both a second method of distinction between prior formats as well as a demonstration of their opposing implications for actual behaviour. This distinction therefore drew on the previously noted differences between prior formats when applying principles of simplicity: while both priors are likely to prefer a lower number of components to simplify the final distribution, this takes two different forms according to the structure of these components, with the Gaussian mixture prior preferring to group more observations together to produce broader components, and the categorical prior limiting the number of values considered in the distribution to only a few key values. It should then be possible to reveal which of these priors is used in this task by encouraging a reduction in components and observing which of these two reactions is displayed: the Gaussian mixture prior should move towards broader components, thereby covering more potential values and so allowing for more varied responses, while the categorical prior should focus on fewer potential responses, most likely limiting a bimodal such as that used in Sanborn and Beierholm (2016) to only the modes of the distribution, essentially turning the task into a high/low classification problem (illustrated in Fig. 1c and d). This could be achieved by introducing uncertainty to the existing design of Sanborn and Beierholm (2016); if the true value of an observation is uncertain, both structures are likely to rely more on their current priors than this new data, encouraging the assignment of that observation to an existing component rather than assuming the presence of a new component.

It should then be possible to identify whether learners are using a truly discrete categorical prior or a continuous Gaussian mixture prior in this case by introducing uncertainty to the dot numeration task of Sanborn and Beierholm (2016) and observing its effect on behaviour. The best method to achieve this is to cause doubt in the feedback given during the task whilst still providing the true value

of the observation: if participants were made to distrust the feedback, for example by stating that this information was accurate in only a subset of trials, participants would no longer be able to rely on the definitive values offered in the original design even where this information was in fact accurate, likely leading to more confusion between actual values based on perceptual data. This allows for the addition of uncertainty to the task without changing any of the specific elements of the stimuli or feedback, instead manipulating uncertainty through instruction alone. What is more, such a manipulation represents a fairly valid scenario; real-world feedback is not always as reliable as that used in laboratory studies, potentially being noisy or vague, or originating from an untrustworthy source. In addition, this design also provides a simple method of manipulating the degree of uncertainty according to the apparent accuracy rate of feedback, allowing for an easy comparison between high and low levels of uncertainty.

The following experiment therefore sought to investigate the processes underlying numerical estimation by adding such an instructional feedback uncertainty manipulation to a numerical judgement task in which participants were trained on a complex distribution through experience. This then provides a contrast of the competing hypotheses of the two potential formats introduced above: if participants are using a categorical prior, responses should be more polarised where feedback is less reliable, focussing mainly on the modes of the distribution. In contrast, if participants are using a Gaussian mixture prior, responses should be more spread out in this case, leading to more midrange and out-of-range responses. This also provided behavioural data for comparison with computational models of the task following these formats for a quantitative suggestion of the underlying process.

2. Experiment 1

2.1. Method

2.1.1. Participants

Forty University of Warwick students were recruited as participants in the experiment from the university's online SONA system in return for £8 in payment. The sample included twenty-five females and fifteen males, while age ranged between 18 and 39 years, with a mean of 22.4.

2.1.2. Design

The experiment used an edited form of the dot estimation task of Sanborn and Beierholm (2016) in which participants were trained on an underlying distribution of dot values through an extensive series of estimation trials: in each trial, a number of dots appeared on the screen for 400 ms, and participants were asked how many they believed had appeared. Dot counts were sampled from a bimodal distribution, ranging between 23 and 32 dots, with modes at the extremes of the range (illustrated in Fig. 1a).

After giving each estimate, a feedback slide appeared noting both the participant's response as well as the true dot count from that trial. In order to induce uncertainty in the feedback, a cover story was used in which the true dot count was given to participants, but presented as a response given by a previous participant for that trial, with the level of uncertainty being manipulated according to the previous participant's reported accuracy rate across all estimation trials. The experiment therefore made use of a between-subjects uncertainty manipulation, using two uncertainty conditions: a high-uncertainty condition, in which the previous participant was stated to be accurate in 70% of trials, and a low-uncertainty condition, in which the accuracy rate was stated to be 95%. This rate was noted on every feedback slide to ensure participants were aware of uncertainty information. Note that while feedback was framed as a response from a previous participant, the stated value was always the true dot count from that trial, providing equivalent information across both conditions.

A discrimination task was also used in the experiment to assess the participant's discrimination ability for use as a parameter in later analysis. In the discrimination task, two sets of dots appeared sequentially on screen, and participants were asked which set (1 or 2) they believed to contain more dots. This was then followed by a feedback slide noting whether the response was correct or incorrect; this was not however affected by the uncertainty manipulation applied to feedback in the estimation task, being definitively accurate in all trials.

2.1.3. Procedure

Upon arriving at the lab, participants were first randomly assigned to one of the two uncertainty conditions, determining the reported rate of accuracy in feedback values. This was balanced to provide equal numbers of participants in each condition, meaning 20 participants were assigned to the high-uncertainty (70%) condition and 20 participants were assigned to the low-uncertainty (95%) condition.

Participants were told the experiment examined how decisions were made under uncertainty, and would involve estimating the number of dots appearing on screen. Participants first performed a set of 128 discrimination trials to assess their initial discrimination ability; this began with a series of 4 practice trials at low dot counts (1–4) to introduce the task.

After this first discrimination block was completed, participants then moved to the estimation task, again beginning with a set of 3 practice trials at low dot counts to introduce the task. Participants performed 500 total estimation trials, with breaks every 50 trials.

Once all estimation trials were completed, participants then performed another round of 128 discrimination trials to track any improvement in discrimination ability. Finally, participants were debriefed as to the aims and expectations of the study.

2.2. Results

Data from one participant was removed from analysis for failing to provide any responses within the presented dot range, leaving

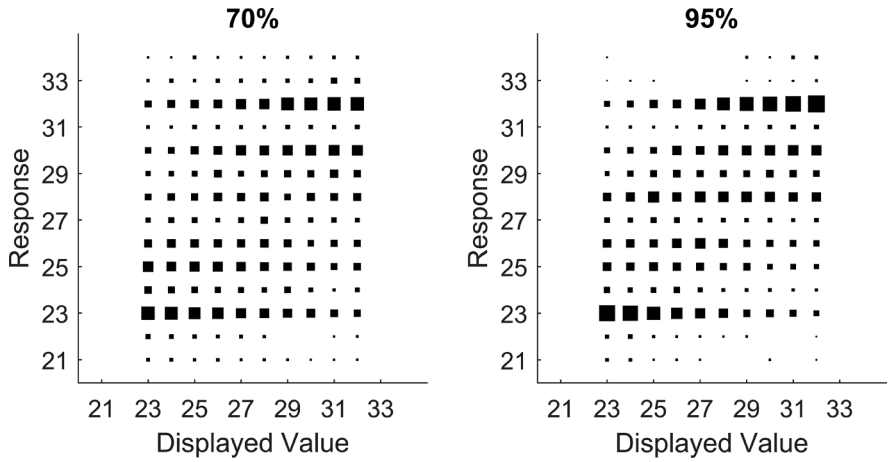

Fig. 2. Conditional response distributions from the 70% and 95% uncertainty conditions of Experiment 1, where square size is proportional to the percentage of responses made to each displayed value.

39 subjects for comparison, with 19 in the 70% condition and 20 in the 95% condition. Responses further than 10 points outside of the displayed range were classified as response errors and removed from analysis; this eliminated an average of 1.81% ([1.40%, 2.27%] 95% confidence interval) of responses across participants.

Fig. 2 shows average response rates for each presented dot value from the two uncertainty conditions; both groups demonstrated reasonable acquisition of the bimodal structure, showing greater preference for the modes of the distribution in their responses. Unshown values were however also used as responses in both conditions, in keeping with the bleed-over predicted by the use of a continuous prior.

The key empirical contrasts from Experiment 1 are summarised in Table 1. Analysis began by contrasting the count of unique responses from the two conditions: this was found to be significantly higher in the 70% group, $t(37) = 2.06$, $p = .047$, $d = 0.69$, with these participants using a wider range of values in their answers. No significant difference was found between the 70% and 95% groups however in either the number of responses from outside the dot range, $t(37) = 1.51$, $p = .140$, $d = 0.51$, or the number of mid-range (non-mode) responses, $t(37) = 0.54$, $p = .590$, $d = 0.18$, though both were found to be higher in the 70% condition.

The data therefore provides some support for the predictions of the continuous mixture prior: while participants in the high-uncertainty condition did not reliably offer a higher number of non-modal responses compared to the low-uncertainty condition, these participants did use a wider range of values in their responses, suggesting the use of a broader set of components when feedback was unreliable. This then provides limited evidence that numeric estimates rely on continuous numerical formats despite the discrete nature of stimuli and responses, utilising the inherent flexibility of such a system to adapt the representation to best capture external data patterns. The lack of reliable differences in all behavioural comparisons does however weaken this conclusion, meaning more substantial evidence is required before this suggestion can be accepted.

In order to address this concern and provide more confidence in the above conclusion, we decided to run a second experiment to further investigate this distinction using the same design but an alternate underlying distribution intended to provide a clearer separation between the two models. This followed the design of the third experiment of Sanborn and Beierholm (2016) in which a more complicated quadrimodal distribution (illustrated in Fig. 3) was used in place of the initial bimodal as a method of further distinguishing between categorical and Gaussian mixture formats: such a distribution is more difficult to emulate using a mixture of continuous components, making the two prior formats more distinct. The use of such a distribution in the present study also provides a clearer separation in empirical measures: the quadrimodal provides a set of values in the middle of the displayed range that are not used in feedback, but may benefit from bleed-over from the two nearby modes under a continuous format. As such, if estimates in this task are in fact based on continuous prior structures, the use of a quadrimodal distribution should offer a clearer demonstration of these effects in both empirical and computational results.

Table 1
Mean measures with bootstrapped 95% confidence intervals from the two uncertainty conditions of Experiment 1.

Condition	70%	95%
Unique response count	17.4 (15.8, 18.9)	15.3 (14.1, 16.5)
Out-of-range responses	54.1 (34.7, 75.4)	30.5 (13.4, 55.0)
Mid-range responses	268 (217, 316)	250 (207, 291)

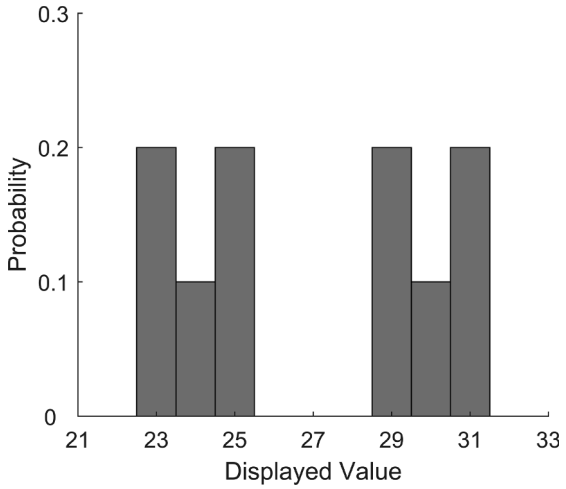

Fig. 3. The quadrimodal distribution used in Experiment 2.

3. Experiment 2

Experiment 2 replicated the dot counting design of Experiment 1 using a more complicated quadrimodal distribution with the aim of providing a stronger contrast between the operation of the discrete and continuous priors. As such, the hypotheses of this experiment were identical to the first, expecting a greater range of responses in the more uncertain condition under a continuous system and a smaller number of responses under a discrete system, though the design was expected to be more diagnostic in separating these hypotheses in this case. In addition, this task also used a larger sample size to provide more statistical power given the reasonably weak findings of the first experiment.

3.1. Method

3.1.1. Participants

Sixty University of Warwick students were recruited as participants in the experiment from the university's online SONA system in return for £6 in payment. The sample included 36 females and 24 males, while age ranged between 18 and 39 years, with a mean of 22.5.

3.1.2. Design

The design of Experiment 2 was identical to that of Experiment 1 with the exception of the underlying distribution: in place of the bimodal distribution, a quadrimodal distribution was used (illustrated in Fig. 3).

3.1.3. Procedure

Experiment 2 used the same procedure as Experiment 1. Assignment to uncertainty conditions was again randomised and controlled to provide equal numbers in each group, meaning 30 participants were assigned to the 70% condition and 30 to the 95% condition.

3.2. Results

Data from Experiment 2 was analysed using the same procedure as Experiment 1, including the same exclusion criteria; while no participants were entirely removed from analysis in this task, an average of 2.33% ([1.71%, 3.10%] 95% confidence interval) of responses across participants fell more than 10 points outside of the displayed range, and so were classified as response errors and eliminated from subsequent comparisons.

Fig. 4 shows average response rates for each displayed dot value from the two uncertainty conditions in Experiment 2; as with the previous experiment, both groups demonstrated reasonable acquisition of the true quadrimodal structure, again showing greater preferences for the modes of the distribution. As in Experiment 1, however, participants in both conditions did also use unseen values in their responses, in this case including the unused values from the midrange of the distribution, again suggesting bleed-over in both groups.

Comparisons from the second experiment are summarised in Table 2. As in Experiment 1, the count of unique responses was found to be significantly higher in the 70% condition, $t(58) = 2.21$, $p = .031$, $d = 0.59$, showing a greater range in the more uncertain condition. Once again, however, no significant difference was found between the 70% and 95% groups in either the number of out-of-range responses, $t(58) = 0.53$, $p = .600$, $d = 0.14$, or the number of mid-range (zero-probability) responses, $t(58) = 0.80$, $p = .425$, $d = 0.21$, though these were again both higher in the 70% group.

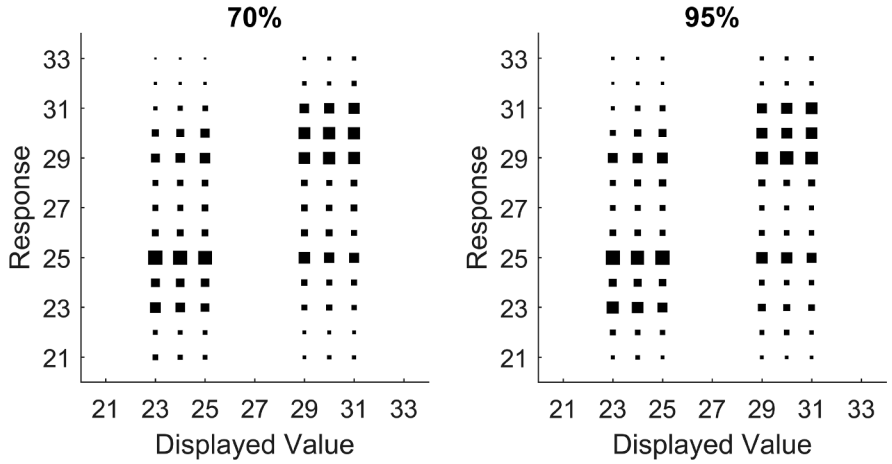

Fig. 4. Conditional response distributions from the 70% and 95% uncertainty conditions of Experiment 2, where square size is proportional to the percentage of responses made to each displayed value.

Table 2
Mean measures with bootstrapped 95% confidence intervals from the two uncertainty conditions of Experiment 2.

Condition	70%	95%
Unique response count	17.0 (15.6, 18.5)	14.7 (13.2, 16.1)
Out-of-range responses	66.3 (44.9, 90.4)	56.2 (31.4, 87.7)
Mid-range responses	69.2 (47.4, 92.7)	55.6 (32.7, 81.1)

These results therefore correspond with the findings of the first experiment: participants in the high-uncertainty condition used a wider range of values in their responses, but did not demonstrate a reliable increase in the use of unshown values over those in the low-uncertainty condition. This again provides limited evidence for the use of a continuous mixture prior, with components seemingly becoming broader under uncertainty, thereby covering more potential values. However, while both experiments may offer weak demonstrations of continuous effects in isolation, these results combine to provide more substantial evidence, suggesting behaviour in these tasks was in fact based on the use of a continuous numerical system.

The collected empirical data then provides a reasonable qualitative indication of the numeric format underlying estimation based on a theoretical contrast of the behaviour of the two considered priors: reactions to uncertainty better match the predictions of a continuous system than a discrete system. To supplement these findings, however, behavioural data was next directly compared with computational models of estimation for a quantitative assessment of the fit of both the continuous and discrete priors to the collected data. This also allowed for an examination of general behavioural trends across all participants beyond the distinction between the two uncertainty conditions of these empirical contrasts, offering an alternate exploration of the processes underlying behaviour in these experiments.

4. The Uncertain Estimation Model

In order to investigate the underlying processes used in the experimental tasks, we developed a perceptual estimation model which was able to use either a continuous or discrete prior format while other model elements remained identical. This drew on existing clustering models in which observations are assigned to subgroups based on similarities in features as well as subgroup size, most notably the Rational Model of Categorisation (RMC) by Anderson (1991) which uses Bayes' rule to approximate the ideal partition of items. As noted in the introduction above, these methods are valuable for their substantial level of flexibility: clustering methods are able to discover patterns in observed data rather than beginning with a pre-set component structure, allowing for much richer representations than parametric alternatives. This is particularly relevant to the present study as pre-defined component structures are unlikely to be able to capture the complexity of distributions such as those trained in these experiments: individual Gaussians cannot adequately match such multimodal structures, while discrete formats require pre-defined ranges that may not be appropriate to all tasks. This flexibility has allowed the application of such systems in previous studies of numerosity (Gershman & Niv, 2013), as well as other topics such as language comprehension (Goldwater, Griffiths, & Johnson, 2009) and causal reasoning (Buchsbaum, Griffiths, Plunkett, Gopnik, & Baldwin, 2015).

The present model therefore considers potential assignments of observations to subgroups based on perceptual data, trial feedback and prior experience in the task, creating a set of clusters which can be aggregated to provide a representation of the true external distribution. The format of these clusters however is dependent on the utilised prior, here limited to the previously noted categorical

and Gaussian mixture priors to contrast discrete and continuous numerical structures. The model is therefore nearly identical to the definitions of the RMC given by Anderson (1991) for discrete and continuous dimensions, here adapted to infer a physical feature for a set of cluster members rather than a category label. It is also notable that the present discrete mixture construction is equivalent to a Dirichlet distribution, as detailed further in Appendix A.6. This model was named the 'Uncertain Estimation Model', or UEM; the following section provides a non-technical description of the operation of this model, while full definitions are available in Appendix A.1.

With each observation, the UEM must determine how to partition the observed items into clusters, calculating the probability of both which cluster each observation will be placed in, and what value that observation will hold. This breaks down into four parts, the combination of which determines this probability: 1. the fit of the perceptual stimulus to each considered value; 2. the fit of the feedback data to each value; 3. the fit of each value to each potential cluster; and 4. the probability of each cluster given its size.

The first of these elements reflects the probability of each potential value producing the observed perceptual stimulus, providing a measure of support for that value from external data; for this purpose, the model uses a lognormal distribution around the displayed value, with variance based on the perceptual precision of the observer.

Similarly, the second element reflects the probability of each potential value producing the given feedback figure, treating feedback information as a perceptual feature of the trial rather than a definitive label. This then allows the model to account for unreliable feedback, assessing the fit of this information to the considered value rather than accepting it as definitive information; for this purpose, the model uses a parameter to reflect the assumed accuracy of the observed feedback figure, with other values dividing the remaining probability. This then means that when feedback is thought to be less reliable, this distribution becomes more uniform, leading to greater reliance on the current prior; as such, observations are more likely to be added to existing clusters than creating new clusters, leading to fewer components overall. Given the supposed social origin of this information, however, this distribution also includes a lognormal noise function around the feedback figure to provide greater support to nearby values; this represents the separate perceptual distribution of the 'past participant' from which feedback is reportedly taken.

The third element then measures the probability of each potential value appearing in each cluster given its membership at that point, as well as a potential new cluster without any members, independent of perceptual or feedback information from that trial. This then introduces the distinction between the previously noted continuous and discrete prior formats: clusters in the categorical prior may contain only one value, meaning any future members must hold the same value as its present members. In contrast, clusters in the Gaussian mixture prior may hold differing values if sufficiently similar, allowing for more variation in future members. The UEM is therefore divided into two subforms according to this difference in cluster format: the discrete UEM (dUEM) and the continuous UEM (cUEM). This also means that the two formats hold distinct hyperpriors, with the continuous cluster format using additional parameters to allow for variation in component width; full detail on the definition of these hyperpriors is given in Appendix A.2.

Finally, the fourth element weights each potential cluster by the size of its membership using a Chinese Restaurant Process (Aldous, 1985; Pitman, 2002) with the inclusion of an additional free parameter to bias the partition towards either large or small clusters.

The combination of these four elements then provides a distribution which defines the probability of both the value and cluster assignment for each observation. These potential partitions can then be aggregated to give a representation of the true external distribution, allowing for predictions of the likelihood of similar values in future trials. The distributions shown in Fig. 1 show an idealised form of this process: in the discrete case, each potential value has a separate component, with the resulting distribution reflecting the proportion of trials assigned to that component as any future members must hold the same value. In the continuous case, future component members may vary in accordance with the variation seen in current members, leading to a wider spread in the centre of the range, but narrower spreads at the modes. Alternatively, by removing the feedback element from this aggregate, the UEM is able to produce a distribution which describes the probability of a response to a particular perceptual stimulus before receiving feedback, matching with the above experimental procedure and so allowing for direct comparison between model predictions and observed behaviour.

To illustrate the predictions of the models, Fig. 5 shows the conditional response distributions of the two models changing only the assumed accuracy of feedback, each averaged across 10 simulations of the task from Experiment 1[1]. These distributions were used to calculate similar measures to those used in the above experimental comparisons: the probabilities of mid-range and out-of-range responses were summed to provide a predicted non-modal response probability, while the response distributions for each trial were used to sample simulated responses to estimate likely unique response counts. In keeping with the hypotheses presented above, non-modal responses become less likely for the dUEM ($p = 0.09$ vs. 0.18) but more likely for the cUEM ($p = 0.50$ vs. 0.41) where feedback is less reliable; the cUEM therefore predicts a higher count of unique responses (18.2 vs. 17.5), whereas the dUEM predicts a lower count (11.2 vs. 11.8), being more likely to rely on the modes of the distribution.

4.1. Model comparison

The discrete and continuous forms of the UEM were compared with the experimental data from both Experiments 1 and 2 using a grid point search across the four parameters shared by the two models. Parameters unique to the cUEM were fixed at predetermined values to make the search computationally tractable, though due to limited manual adjustments to decide these values on a subset of

[1] Parameters for this simulation were: $c = 0.7$, $e = 1$, $w_b = 0.01$.

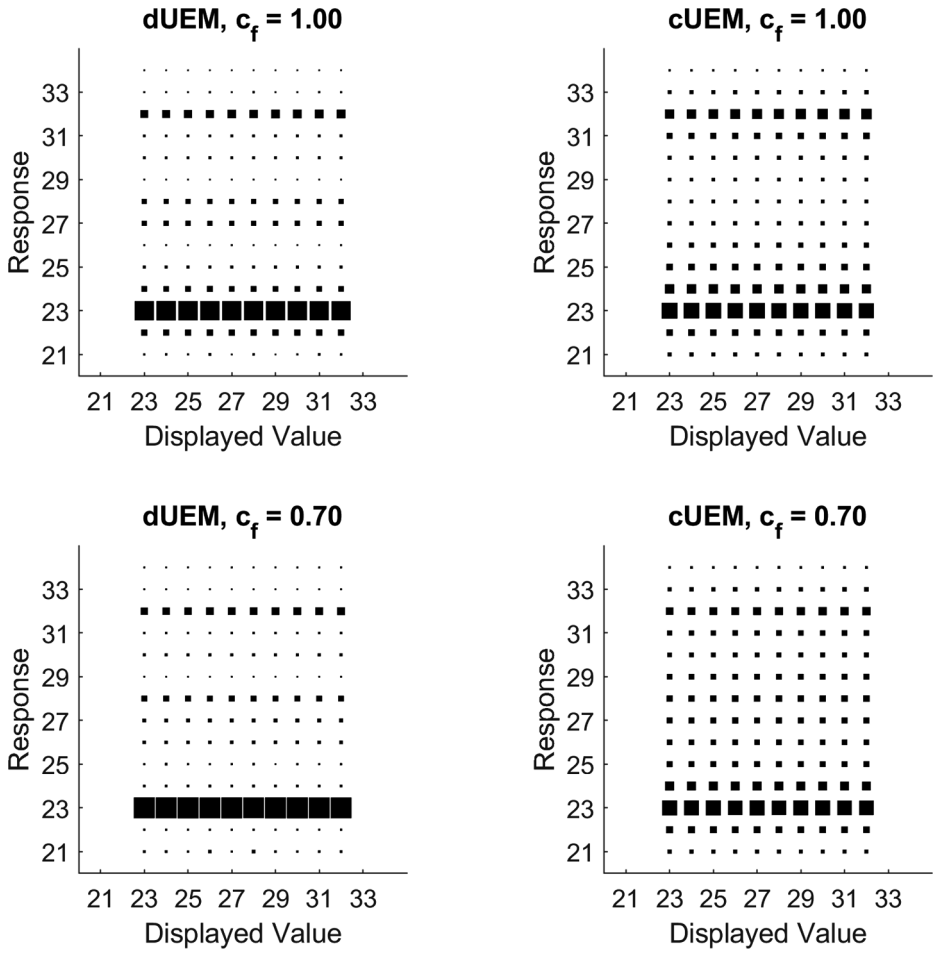

Fig. 5. Simulated conditional response distributions from the dUEM and cUEM where all parameters are fixed except for assumed feedback accuracy; the top figures assume accuracy is absolutely correct, while the lower figures assume accuracy is 70% correct.

the data, these were treated as manipulated. As such, the dUEM was defined as having four free parameters, and the cUEM was defined as having six. In addition, due to stochasticity in the clustering process of both models, each grid point was repeated 10 times to produce an average likelihood estimate for that set of parameters. Full details of this procedure are given in Appendix A.2.

Both models were fit to each participant individually to provide maximum likelihood values for each model for each participant, which were then converted to Akaike information criterion (AIC, (Akaike, 1974)) and Bayesian information criterion (BIC, (Schwarz, 1978)) values for further comparison due to the differing number of parameters between the models. These measures both provide an adjusted measure of model fit controlling for model complexity, with lower values indicating a better fit. For ease of presentation, the following analysis focuses primarily on BIC scores as the more conservative measure, with AIC results being noted where these scores suggest a qualitative difference in outcome, while full AIC results are listed in Appendix A.3. Both measures were also used to calculate weights for the given comparison between the cUEM and dUEM, providing an estimate of the posterior probability of each model assuming equal priors (Wagenmakers & Farrell, 2004). To provide a measure of the variation in likelihood estimates arising from model stochasticity, bootstrapped 95% confidence intervals were calculated for the key model measures by resampling the obtained likelihoods at each grid point to provide new average likelihood estimates and recalculating maximum likelihood, AIC and BIC values for each participant; this was repeated for 10000 iterations, with intervals being taken from the quantiles of these distributions. This also allows for confidence intervals on the number of participants best fit by each model, as reported below.

The grid point structure allows for both global and individual fitting, either aggregating likelihoods across participants at each grid point assuming a common set of parameters within each experiment, or calculating a maximum likelihood score for each participant assuming differences in parameters and then aggregating the resulting BIC scores. Both AIC and BIC measures from both experiments do however show fits are substantially better when using individual parameters; as such, the remainder of the comparison uses aggregates of individual best fits, with global fits being listed in Appendix A.3.

Fig. 6 shows the aggregated conditional distributions from the best fitting parameters for Experiment 1, while full quantitative results are given in Table 3. Across all participants, the cUEM had a better fit to the data by summed BIC scores than the dUEM. On an individual basis, a large majority of participants were better fit individually by the cUEM (33 [30–36 95% CI]), with a small

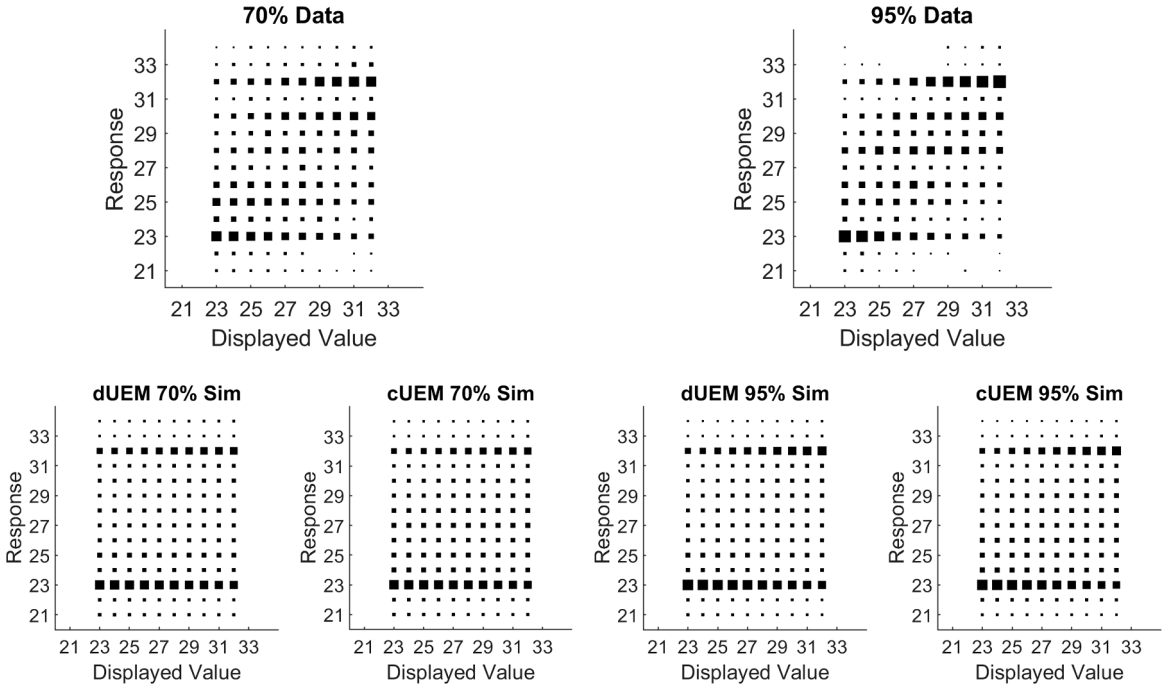

Fig. 6. Averaged conditional response distributions from the maximum likelihood estimates of the discrete and continuous models in Experiment 1, separated by uncertainty condition, including empirical data for comparison.

Table 3
Modelling results from Experiments 1 and 2, reporting the best fitting model for each comparison between the discrete and continuous models and the margin of this advantage in summed maximum log likelihood across participants for that model (MLL), AIC and BIC scores. *w*(AIC) and *w*(BIC) are the weight of the AIC and BIC scores respectively for the given comparison, while brackets provide bootstrapped 95% CIs, omitted for weight measures as these are 1 in all cases.

Experiment	Comparison	Best Model	ΔMLL	ΔAIC	*w*(AIC)	ΔBIC	*w*(BIC)
Experiment 1	Individual	cUEM	1855 (2161, 1691)	−3554 (−3225, −4166)	1	−3225 (−2897, −3837)	1
	70%	cUEM	1109 (1352, 953)	−2358 (−1830, −2627)	1	−1980 (−1669, −2467)	1
	95%	cUEM	746 (906, 658)	−1413 (−1235, −1731)	1	−1244 (−1067, −1562)	1
Experiment 2	Individual	cUEM	2677 (3085, 2480)	−5115 (−4720, −5930)	1	−4609 (−4214, −5425)	1
	70%	cUEM	1389 (1670, 1254)	−2658 (−2388, −3220)	1	−2405 (−2135, −2967)	1
	95%	cUEM	1289 (1537, 1102)	−2457 (−2085, −2953)	1	−2205 (−1832, −2700)	1

proportion being better fit by the dUEM (6 [3–9]). When separated by uncertainty condition, the cUEM provided a better fit to both groups, accounting for 15 (14–18) of the 19 participants in the 70% group and 18 (16–19) of the 20 participants in the 95% group. To investigate the suggestion that behaviour appears more discrete where feedback is more reliable as this is where continuous components are best able to emulate discrete structures, a chi-squared test compared the ratio of participants best fit by the two models between uncertainty conditions, finding no significant difference, $\chi^2(1) = 0.26$, p = .608.

AIC scores offer almost identical qualitative results, both in the fit of the models across participants and the proportions best fit by each model, though margins between AIC scores demonstrate a stronger support for the cUEM, showing greater differences across all comparisons. Such results are attributable to the reduced cost of complexity in AIC scores, more closely reflecting the difference in raw likelihood despite the different parameter counts of the two models. This is notable given that the current comparisons did not take full advantage of the greater complexity of the cUEM, as the additional parameters of this model were in fact fixed across the comparison, but were treated as variable given the initial manual manipulations of variance and confidence to allow for narrower components. This does however mean that the cUEM performed better even under the harsher complexity costs of the BIC measures, providing further support for this prior.

Model fitting results from Experiment 2 are illustrated in Fig. 7, while quantitative measures are listed in Table 3. As with the first experiment, the cUEM displayed an advantage in the number of participants best fit by each of the models, accounting for 47 (44–51) of the 60 participants; this is further displayed in the summed BIC scores, which again show the cUEM had a better overall fit to the data. Separated by group, summed BIC scores again found the cUEM to have a better fit in both conditions, accounting for 25 (23–27) of the 30 participants in the 70% condition, and 22 (20–25) of the 30 participants in the 95% condition. As with the first experiment, the difference in model ratios between the two groups was found to be non-significant, $\chi^2(1) = 0.39$, p = .531.

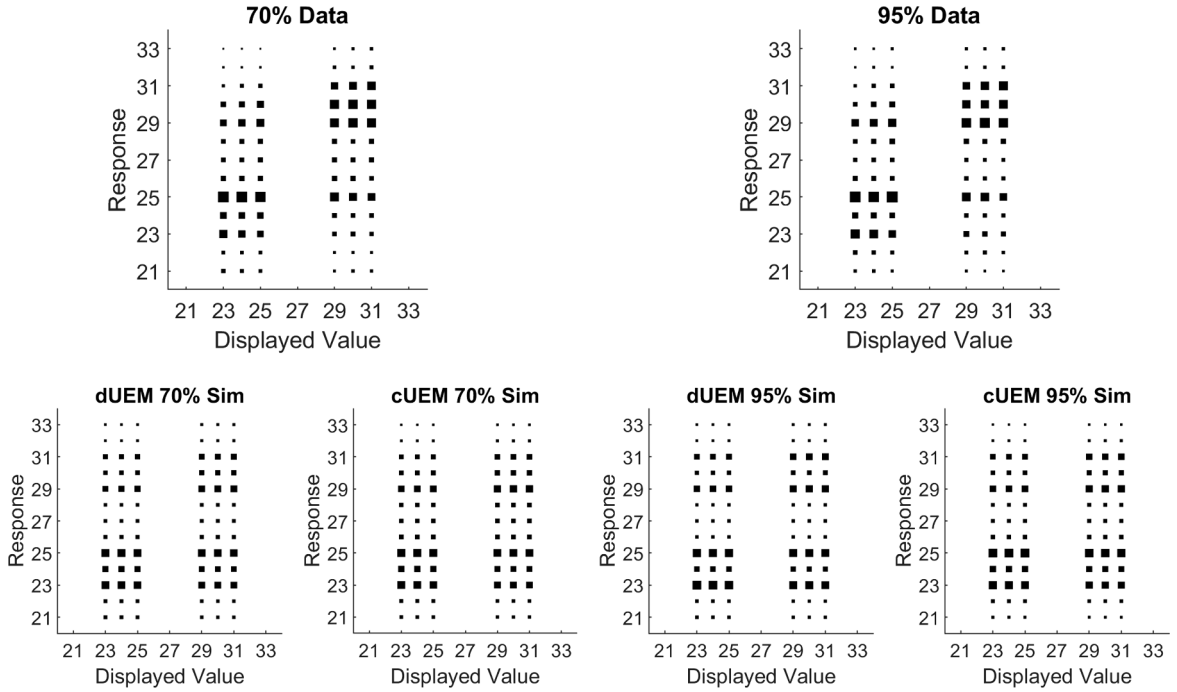

Fig. 7. Averaged conditional response distributions from the maximum likelihood estimates of the discrete and continuous models in Experiment 2, separated by uncertainty condition, including empirical data for comparison.

AIC scores meanwhile again show almost identical results, though with slight differences in the ratios of participants best fit by each model, again seemingly showing greater support for the cUEM where the penalty for complexity is less severe.

Results from both model comparisons therefore suggest that a Gaussian mixture prior was more likely to be used in their respective tasks than a categorical prior, so supporting the apparent continuous effects observed in the empirical contrasts. These comparisons then correspond with the above empirical findings: while behavioural data in both experiments demonstrates qualitative evidence of a continuous representation of past numeric experience, this is now reinforced quantitatively by model fitting, providing greater confidence in this conclusion. This highlights the difference between the empirical and computational comparisons used here: while empirical contrasts focus on the differences in behaviour between the two uncertainty conditions, which may be limited in scope, the model comparison is able to examine wider behavioural patterns across all participants, identifying a trend towards continuous behaviour common to both groups.

This conclusion does however rely on the assumption that all participants use a common model, which may be questionable given the division between model fits observed at the individual level: a number of participants in both tasks were better fit individually by the dUEM, suggesting the continuous model does not provide the better description for all participants. As such, while the continuous prior offers a strong fit across behaviour collectively, this may not be a truly universal system, with potential individual differences in prior format between participants. To further examine these differences, an additional model selection analysis was performed following the procedure outlined by Stephan, Penny, Daunizeau, Moran, and Friston (2009) and Rigoux, Stephan, Friston, and Daunizeau (2014). This analysis treats the model as a random effect between subjects following an underlying distribution across the population, providing estimates of both the broader frequency of each model, and the 'protected exceedance probability' that a given model accounts for a greater proportion of subjects than other candidates. Again, bootstrapping was used to provide 95% confidence intervals on these measures to account for stochasticity in the model fits. Results from this analysis corresponded with the division found between model fits at the individual level reported above: estimated model proportions were higher for the cUEM (0.826 [0.785–0.872]) than the dUEM (0.174 [0.129–0.215]), while the protected exceedance probability of the continuous model was extremely high ($p > .999$ in all cases), suggesting the continuous model was significantly more common in the participant sample. This further reinforces the suggestion that the use of continuous numerical formats is dominant within our sample, though a minority may instead use a discrete system.

There is however an important caveat to the above results: the stochastic clustering processes used by both models introduce substantial variation in likelihood estimates even with identical parameter values, as repeated simulations can produce different predicted partitions of observations, and therefore different response distributions. This is partially demonstrated by the sizeable confidence intervals for the model fitting measures reported above, but can be observed directly in the standard deviation in likelihood at the best fitting points from the above exercise: average standard deviation across all participants was 199.86 for the dUEM and 213.36 for the cUEM. As such, any individual fits from this comparison should be taken with caution, as current likelihood values may not allow sufficient precision to characterise the methods used by specific learners. Even so, the results do provide reasonable

confidence in the reported group-level effects, with the continuous model providing the better fit to the participant sample on aggregate by a wide margin.

It should also be noted that these results focus purely on a direct contrast of the two candidate models rather than the absolute fit of these models to the data; while one model might outperform another in relative terms, this does not reveal whether either model offers an accurate account of behaviour more generally. To provide a measure of absolute model fit, correlations were calculated between the conditional response distributions of each participant and those generated from the maximum likelihood estimate of each model to that participant's data, illustrated in Figs. 6 and 7. These correlations showed moderate results (mean R^2: dUEM = 0.434; cUEM = 0.453), suggesting that these models alone may not provide a complete account of behaviour in these tasks. This is displayed visually in Figs. 6 and 7: both models do capture the preference for the modes of the respective distributions in responses, but also seem to make greater use of mid-range values than was demonstrated by participants. The present models may then require further development to fully capture the process of human estimation; even so, these definitions should satisfy the comparison of continuous and discrete numerical formats which remains the main focus of this study. It is also notable that these measures show a higher average fit for the continuous model than the discrete model, supporting the results of the direct model comparison, though this is a less sensitive measure of relative fit than the BIC values used above.

Finally, as an additional test of the discriminability of the two models, a model recovery exercise was performed in which sets of simulated responses were generated from each model and then fit by the two candidate models to examine whether the fitting procedure is able to accurately identify the true generating process. 100 sets of simulated data were generated for each of the two models using the best fitting parameters found for the 100 collected participants from Experiments 1 and 2 in the above model comparison. The models were then fit to the simulated data using the same procedure as the participant data, as detailed in Appendix A.2, determining the best fitting model for each simulated subject. Model recovery rates were then calculated by taking the proportion of simulated subjects created for each model that were best fit by their respective generating model; these rates were reasonably high for both models, though slightly higher for the continuous system (dUEM: 0.79; cUEM: 0.87). This suggests the models are fairly discriminable, though as previously suggested, the continuous model may be better able to mimic the discrete model than vice versa, seemingly being able to accurately capture data generated through discrete systems. It should be noted however that these recovery rates are based on data directly generated by the candidate models, whereas discriminability based on actual participant data is less certain, as discussed above.

5. Discussion

The above sections provide evidence from two experiments of a continuous numerical system underlying discrete estimates which reacts to uncertainty by simplifying the held representation according to rational categorisation principles: in both tasks, responses became more varied when feedback was less reliable, indicating a broadening of Gaussian components. This is further supported by comparisons with computational models of estimation: in both experiments, behaviour was better fit by a Gaussian mixture prior over a discrete mixture prior in both the number of participants accounted for and aggregated measures of fit, providing a second source of evidence for the use of a continuous prior format. This conclusion is not completely definitive, however: empirical data do not show reliable continuous effects in all measures, while noise in computational data adds ambiguity to individual-level fits. Even so, these results do indicate a common tendency towards the use of continuous numerical formats at the aggregate level, finding greater support for continuous effects across subjects even in a scenario in which targets, responses and feedback were all discrete.

Such findings offer an interesting contrast with the findings of Sanborn and Beierholm (2016), where a large majority of participants were better fit by discrete systems. While this could be attributable to a difference between the participant samples used in these studies, this may instead demonstrate the benefit of fitting full continuous mixture models to behaviour: these direct comparisons offer a more sensitive analysis of behaviour, potentially revealing that continuous systems are more common than such results would suggest. This follows the suggested emulation of discrete structures by continuous systems described in the introduction to this study: learners may be able to acquire complex multimodal distributions through the use of a highly flexible continuous numerical system able to emulate such detailed structures. This then allows for the appearance of the use of discrete numerical formats in such tasks despite actually being based in continuous systems, offering a new framing of the results of Sanborn and Beierholm (2016): the apparent use of discrete priors in that study may in fact be the result of a continuous system emulating the narrower component format of a truly discrete distribution. This may be attributable to aspects of the design of that study which facilitated such emulation: for example, the range of values displayed in the task was reasonably small in comparison to other studies (e.g. Gershman & Niv, 2013), potentially encouraging the use of a set of narrow components to provide better discrimination. Alternatively, the use of definitive rather than unreliable feedback may have avoided potential noise in value assignment which could broaden components: without any reason to believe feedback is inaccurate, participants may have been able to further narrow their components for a closer emulation of discrete structures. Further work will therefore be required to fully investigate the prevalence of these systems across learners, and whether the continuous preference observed here is similarly displayed in other tasks and populations.

These findings do however correspond with the suggestions from previous research noted in the introduction to this study that learners have both a continuous approximate number system and a discrete symbolic number system available to them when constructing numerical representations through experience (Dehaene, 2011). The present results then offer an interesting display of the use of continuous systems even in discrete numerosity judgements: use of a continuous numerical format appears dominant in the present task despite stimuli, responses and feedback all being discrete. This in fact corresponds with previous numerical research in which numbers often appear to be considered within a continuous format: even when presented symbolically, behaviour seems to

suggest numerical values are treated continuously, showing greater confusion between similar values (Moyer & Landauer, 1967; Spelke & Tsivkin, 2001; Dehaene & Marques, 2002). The present study may then further contribute to the suggestion that learners often rely on approximate number systems when dealing with numerical values, translating the output of such systems into discrete figures when required (Izard & Dehaene, 2008). This links to the concept of 'number sense' (Dehaene, 2011), an innate understanding of numerosity displayed independently of the standard symbolic numerical system, as evidenced by its use by not just adult learners, but also infants (McCrink & Wynn, 2004) and animals (Flombaum, Junge, & Hauser, 2005; Ditz & Nieder, 2016).

The apparent use of continuous structures across numerical tasks may then reflect a common preference for this number sense, utilising a more fundamental numerical system where possible and converting this to symbolic formats as needed rather than directly working in a purely symbolic format learned in later life. What is more, the current results demonstrate that despite being a more primitive system, these structures can still enable efficient learning under the right circumstances: within the framework of a rational clustering process, continuous structures can be used to represent reasonably complex distributions, particularly where their inherent flexibility can be exploited. This being said, such a reliance on continuous numerical formats may not be universal, as a small subset of the participants in these experiments were better fit by the discrete model. This could suggest that a minority of learners do in fact prefer to use the symbolic number system to represent their experience with discrete estimates, possibly due to its correspondence with task elements, or potentially for a greater level of precision than is provided by the approximate system. As previously noted, however, the level of noise in likelihood estimates in the present results introduces some doubt to individual model fits, preventing any firm conclusions regarding actual usage rates of these formats. This reinforces the need for further testing on the prevalence of the use of these numerical systems, and whether any observed differences are driven by individual preference or task demands.

It is also notable that the present findings indicate the use of a highly flexible estimation system in which any formed representation and resulting behaviour are highly sensitive to the scenarios that produce them. This applies to both the availability of two number systems, but also the flexibility of these systems themselves: both the continuous and discrete prior formats are able to adjust their structures to suit external data, though the continuous prior does have greater flexibility in this regard given its ability to adjust the variance of its components. Such flexibility allows either system to acquire more complex distributions such as those used in the present experiments: without such a representation, learners would not be able to accurately capture such forms. This can in fact be demonstrated by lesioning the present models to remove their respective priors, thereby basing decisions solely on perceptual evidence; this generates a drop in estimated accuracy for both discrete (15.9% vs. 5.00%) and continuous (16.4% vs. 5.20%) formats, illustrating the benefits to learning provided by such a system (more detail on this procedure is given in Appendix A.5). This flexibility also allows the learner to account for uncertainty in the formed representation, further altering mental structures according to noise in the environment such as the unreliable feedback of the present designs. In addition, recent work has also suggested such systems could offer an advantage in terms of cognitive economy, reducing complex distributions to sets of summary statistics for component clusters to aid representation (Sun, Li, & Zhang, 2019). As such, these results help to demonstrate the power of a rational system in this task, utilising both direct observations and background knowledge to build a mental representation which accurately captures both external patterns and their surrounding context.

In addition to the format of numerical information, the present distinction between discrete and continuous systems also demonstrates the impact of this structure on behaviour through the application of simplicity: the two priors provide almost directly opposing reactions to uncertainty, with one reducing the number of considered responses in order to simplify response selection, and one reducing the number of response regions but allowing for more potential values. The use of these numerical structures therefore carries distinct behavioural implications: use of a continuous system will likely lead to greater reliance on prior expectations where feedback is judged to be unreliable, drawing estimates towards previously expected values, but without necessarily disregarding such information. Returning again to the example of counting people in a room, if the observer receives a potential count from another individual that is viewed as unreliable, under a continuous format they are unlikely to store that figure in memory, but may use a similar number that falls between the feedback figure and their own prior expectations. In contrast, use of a discrete system may show more extreme behaviour, potentially completely abandoning unreliable feedback in favour of prior expectations. Such a distinction is important given that real-world estimates are rarely followed by definitive feedback; even where such information is provided, this can be vague, or from an untrustworthy source. This illustrates the broader importance of understanding the form of our representations, as slight differences in structure can have substantial effects on behaviour. As such, any interventions into such systems must consider what structures people may hold in order to provide meaningful results; in the current case, this applies primarily to methods that may encourage more accurate learning of real-world distributions, though this concept applies to any action based on internal mental representations.

There are however some additional elements to consider regarding these conclusions, beginning with the limitations of the present analysis: as noted above in the results of the model comparison, there is a substantial amount of noise in the estimation of likelihoods for these models, leading to some doubt in model fits for individual subjects. We have therefore focused here on broader group-level effects where support for the continuous model is more assured rather than the relative prevalence of the two models in this sample. Such individual differences in the use of numerical formats do however remain an important consideration, requiring further testing to provide more definitive results. Moreover, measures of absolute fit for the current models are fairly low, indicating neither the discrete nor continuous model definitions used here offers a complete account of learning in this task. These definitions were chosen to provide the clearest contrast of continuous and discrete numerical formats, mirroring the pre-existing distinction between the approximate and symbolic systems given by past research. Given the apparent limitations of these definitions, however, these models will require further development if they are to account for actual behaviour. Future work is therefore clearly required in order to fully understand the processes used in such estimation tasks, both expanding the present models and proposing new potential systems for comparison with behaviour.

This also raises the possibility of considering further learning systems outside of the strict dichotomy between continuous and discrete formats which was the focus of this study: alternate models could in fact bridge these two formats, either switching between systems according to task demands or mixing the two priors to form a hybrid distribution. Such a combination could result in a highly flexible estimation system able to produce either of the behaviours associated with the individual priors described here. This would however come at the cost of significant complexity, not only aggregating the demands of both the priors considered in this study, but also requiring additional learning of the points at which each prior is beneficial if this system is to be effective. Even so, these combinations do remain an interesting possibility, and therefore may need to be considered in future work.

It should also be noted again that the present Bayesian models were used as descriptions of behaviour to facilitate the comparison between discrete and continuous prior formats, and do not necessarily reflect the processes used by actual learners when making numeric estimates. This also places the current models at the computational level of analysis (Marr, 1982), offering high-level principles for behaviour rather than any specific algorithmic mechanism that may be used by actual learners. Even so, BDT does remain a strong candidate for the true process: as previously noted, BDT provides a better account for the use of prior information than theories such as calibration (Sanborn & Beierholm, 2016), allowing for the acquisition of more complex distributions such as those used in the present study. In addition, existing work has offered a number of algorithms which could support Bayesian models such as these, most notably sampling methods (Gelman et al., 2013), which have been found to accurately account for human biases in a number of tasks (Sanborn, Griffiths, & Navarro, 2010; Griffiths, Vul, & Sanborn, 2012; Sanborn & Chater, 2016). The current results are not however able to definitively determine the validity of the considered Bayesian models, meaning these models remain descriptions until more direct tests are performed.

Finally, one additional factor to consider in this study is the method by which uncertainty was manipulated in this design: in order to create doubt in the task feedback, true values were presented as answers given by a past participant, using that participant's reported accuracy rate as a measure of reliability. This therefore introduces a social information element to the task, as participants are made to consider the method by which these feedback values are generated. This is particularly notable given that previous research has found that learners may draw different inferences from observed data according to its origin: beliefs may differ when examples are chosen by a teacher to illustrate an idea (Shafto, Goodman, & Griffiths, 2014), or when samples are noted to be exclude certain results (Hayes, Banner, & Navarro, 2017) compared to observation alone.

While the current task is unlikely to have encouraged these particular higher level inferences, the origin of feedback remains a consideration when determining how participants interpret this information during decision making: there are multiple potential methods of using feedback data with varying levels of complexity, ranging from a reasonably simplistic correct/incorrect dichotomy to a full model of the past participant's decision process. For the purposes of simplifying model fitting, a reasonably basic form of this process was used in both of the present models, using a single parameter to reflect the probability of the feedback being accurate with surrounding noise; future work on this subject may therefore wish to consider these alternate definitions in order to provide a more complete model of behaviour. Alternatively, similar tasks could make use of non-social manipulations of uncertainty to assess the impact of this factor on decision making.

5.1. Conclusion

The present study provides both empirical and computational evidence that discrete numeric estimates are built on continuous mental structures, displayed here via reactions to uncertainty: learners react to unreliable feedback by broadening their response regions, utilising the inherent flexibility of their representation to account for noise in the environment. This demonstrates not just the systems used within numerical estimation, but also the impact of these systems on both the distributions learned through this process as well as behaviour built on this representation. These findings are however limited by uncertainty regarding potential differences in the use of these systems between individuals, requiring further testing to determine the true prevalence of the use of continuous formats in the wider population. We therefore hope that this study can provide a basis for further examination of the mechanisms underlying numerical estimation, using additional experimental contrasts and more advanced computational models to offer greater insight into these systems, and so the wider representation of numerical information.

Declaration of Competing Interest

None.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. JS and ANS were supported by a European Research Council consolidator grant (817492-SAMPLING). ANS was supported by the Economic and Social Research Council (Grant No.: ES/K004948/1). URB was supported by the Leverhulme Trust (Grant No.: RPG-2017-097).

This research has not been previously published in any other form, and is not under consideration in any other publication.

All empirical and computational data collected for this study, as well as the files used in the described analysis and modelling, are available on the Open Science Framework at: https://osf.io/bry3d/?view_only=2c1f6bf89c0c4b65a185eace5524c96f.

Appendix A. Appendices

A.1. Model definition

The following provides the full definition of both the discrete and continuous forms of the Uncertain Estimations Model. On each estimation trial, the model determines the probability of each potential value in each potential cluster generating both the observed perceptual data and the given feedback value across all possible partitions of past observations:

$$p(S_t|X_{1:t}, F_{1:t}) = \sum_{S_{1:t-1}} \sum_{Z_{1:t}} p(S_t, S_{1:t-1}, Z_t, Z_{1:t-1}|X_{1:t}, F_{1:t}) \tag{1}$$

where t is the current trial, $S_{1:t-1}$ is a vector containing the dot counts $S_1, S_2, ..., S_{t-1}$, $Z_{1:t}$ is a vector containing the cluster indices $Z_1, Z_2, ..., Z_t$, $X_{1:t}$ is a vector containing the perceptual data $X_1, X_2, ..., X_t$ and $F_{1:t}$ is a vector containing the feedback values $F_1, F_2, ..., F_t$. This can be broken down to isolate the probability of the proposed value generating the observed perceptual and feedback data:

$$p(S_t|X_{1:t}, F_{1:t}) \propto \sum_{S_{1:t-1}} \sum_{Z_{1:t}} p(X_t|S_t)p(F_t|S_t)p(S_t|S_{1:t-1}, Z_{1:t})p(Z_t|Z_{1:t-1})p(S_{1:t-1}, Z_{1:t-1}|X_{1:t-1}, F_{1:t-1}) \tag{2}$$

This equation is composed of five elements to be calculated: first, $p(X_t|S_t)$ notes the probability of the observed perceptual stimulus given the potential value S_t, where X_t is an estimate of the perceptual stimulus sampled from a lognormal distribution with mean equal to the logarithm of the true dot count v_t and fixed variance σ_l^2 based on assessment of the observer's discrimination ability:

$$p(X_t|v_t) = logN(\ln(v_t), \sigma_l^2) \tag{3}$$

This estimate is then compared with each considered value using a second lognormal distribution with mean equal to the logarithm of the considered value and equal variance:

$$p(X_t|S_t) = logN(\ln(S_t), \sigma_l^2) \tag{4}$$

Secondly, $p(F_t|S_t)$ notes the probability of the feedback score given the proposed value, allowing for the consideration of uncertainty in feedback information. For the purposes of simplicity, this uses a single parameter to reflect the assumed reliability of trial feedback, with remaining probability being spread uniformly over other potential values. Given the supposed social nature of feedback information, however, there may be an assumption that even if inaccurate, the feedback figure should be close to the true value, meaning uniform noise may be invalid. To address this concern, a log-normal noise function was added to the feedback distribution, corresponding with the perceptual distribution given in Eq. (4), before being renormalised:

$$p(F_t|S_t) = \frac{c_f\delta(F_t - S_t) + \frac{1-c_f}{n_v-1}(1 - \delta(F_t - S_t)) + logN(ln(S_t), \sigma_l^2)}{1 + \sum logN(ln(S_t), \sigma_l^2)} \tag{5}$$

where c_f is the feedback accuracy parameter, fixed across all trials, n_v is the number of values considered for S_t, and δ is a Dirac function comparing the proposed value S_t with the feedback value F_t, being 1 where these values are equal and 0 elsewhere. This then assumes that the observer treats the feedback figure as a sample from a perceptual distribution identical to their own, avoiding any substantial modelling of the 'past participant'.

Thirdly, $p(S_t|S_{1:t-1}, Z_{1:t})$ notes the probability of the proposed value given the partition suggested by $S_{1:t-1}$ and $Z_{1:t-1}$ and the proposed cluster Z_t. This term therefore introduces the distinction between continuous and discrete structures, as this affects the generated partition.

A.1.1. Discrete Format

For the discrete form, a count of matching observations is used:

$$p(S_t|S_{1:t-1}, Z_{1:t}) = \frac{n_s}{n_z} \tag{6}$$

where n_s is the count of observations in cluster Z_t with value S_t and n_z is the total membership of cluster Z_t; this distribution therefore becomes binary for non-empty clusters due to the uniformity of their membership, being 1 where S_t matches the value of these members and 0 elsewhere. For new potential clusters without any members, this instead uses a uniform prior across the considered values of S_t. This distribution therefore matches the definition used by the RMC for likelihood values using discrete dimensions where the prior expectancy parameter used by the RMC (α) approaches zero.

A.1.2. Continuous Format

For the continuous form, a Gaussian mixture is used, computing the mean and variance of the cluster distribution given its currently assigned members as well as an assumed prior mean and variance independent of any observations. This follows the definition given by the RMC for likelihoods using continuous dimensions, in which an inverse chi-squared distribution is used to provide an estimate of the variance:

$$\sigma^2 \sim \beta_0\sigma_0^2\chi^2_{\beta_0} \tag{7}$$

where σ_0^2 is the prior variance and β_0 refers to the confidence in this prior variance, while the mean uses a Gaussian distribution:

$$\mu|\sigma \sim N\left(\mu_0, \frac{\sigma}{\sqrt{\lambda_0}}\right) \tag{8}$$

where μ_0 is the prior mean and λ_0 is the confidence in this prior mean (note that the second parameter of this distribution is the standard deviation rather than the variance). The use of these two distributions then results in a t-distribution describing the probability of value S_t in the given cluster (again, the second parameter of this t-distribution is the standard deviation rather than the variance):

$$p(S_t|S_{1:t-1}, Z_{1:t}) = t_c(\mu_i, \sigma_i\sqrt{1 + 1/\lambda_i}) \tag{9}$$

normalised within each cluster to prevent probability exceeding 1 where variance is low. The parameters of this distribution are calculated according to the proposed membership of the target cluster in the currently assumed partition, combining the prior mean μ_0 and variance σ_0^2 with the observed mean $\bar{x}$ and variance s^2 using the confidence values β_0 and λ_0:

$$\beta_i = \beta_0 + n_z \tag{10}$$

$$\lambda_i = \lambda_0 + n_z \tag{11}$$

$$\mu_i = \frac{\lambda_0\mu_0 + n_z\bar{x}}{\lambda_0 + n_z} \tag{12}$$

$$\sigma_i^2 = \frac{\beta_0\sigma_0^2 + (n_z - 1)s^2 + \frac{\lambda_0 n_z}{\lambda_0 + n_z}(\mu_0 - \bar{x})^2}{\beta_0 + n_z} \tag{13}$$

Fourthly, $p(Z_t|Z_{1:t-1})$ is a Chinese Restaurant prior (Aldous, 1985; Pitman, 2002) describing the probability of the observation being assigned to cluster Z_t based on the size of that cluster, following the format of Anderson (1991):

$$p(Z_t|Z_{1:t-1}) = \begin{cases} \frac{cn_z}{(1-c)+cn} & \text{if } Z_t \text{ is old} \\ \frac{(1-c)}{(1-c)+cn} & \text{if } Z_t \text{ is new} \end{cases} \tag{14}$$

where n_z is the number of observations in cluster Z_t in the current partition, n is the total number of assigned observations and c is a coupling parameter describing the probability of two items being grouped together independent of any other observations.

Finally, $p(S_{1:t-1}, Z_{1:t-1}|X_{1:t-1}, F_{1:t-1})$ describes the probability of the currently assumed partition given by $S_{1:t-1}$ and $Z_{1:t-1}$, which is equal to the product of the probability of each past observation's assignment to the partition as defined by Eq. (2).

Once the probability of each potential permutation has been calculated, these values can be used to generate the predictive probability of any value appearing in the next trial by averaging over the individual distributions of each potential partition:

$$p(S_{t+1}|X_{1:t}, F_{1:t}) = \sum_{S_{1:t}}\sum_{Z_{1:t+1}} p(S_{t+1}, S_{1:t}, Z_{t+1}, Z_{1:t}|X_{1:t}, F_{1:t}) \tag{15}$$

$$\propto \sum_{S_{1:t}}\sum_{Z_{1:t+1}} p(S_{t+1}|S_{1:t}, Z_{1:t+1})p(Z_{t+1}|Z_{1:t})p(S_{1:t}, Z_{1:t}|X_{1:t}, F_{1:t}) \tag{16}$$

Similarly, the UEM is able to calculate the probability of the responses made by participants in the present experimental procedure, where estimates are given to the perceptual stimulus prior to receiving feedback, by simply omitting the feedback element from Eq. (2):

$$p(S_t|X_{1:t}, F_{1:t-1}) = \sum_{S_{1:t-1}}\sum_{Z_{1:t}} p(S_t, S_{1:t-1}, Z_t, Z_{1:t-1}|X_{1:t}, F_{1:t-1}) \tag{17}$$

$$\propto \sum_{S_{1:t-1}}\sum_{Z_{1:t}} p(X_t|S_t)p(S_t|S_{1:t-1}, Z_{1:t})p(Z_t|Z_{1:t-1})p(S_{1:t-1}, Z_{1:t-1}|X_{1:t-1}, F_{1:t-1}) \tag{18}$$

A.1.3. Details of model approximations

While the above equations do provide a calculable formula, by considering all possible permutations of past cluster and value assignments, the full version of the model would quickly become intractable at even a moderate number of observations. As such, this full solution is approximated by reducing the number of considered permutations to a set of samples using particle filtering. This process makes use of a fixed number of 'particles', each containing a possible permutation of cluster and value assignments for past trials at that point in time (Griffiths, Sanborn, Canini, Navarro, & Tenenbaum, 2011). Following a new observation, the model considers only the assignments of that observation which are consistent with current particles, calculating the probability of the assignment according to:

$$p(S_t, Z_t|X_{1:t}, F_{1:t}) \approx \frac{1}{n_i}\sum_i p(X_t|S_t^{(i)})p(F_t|S_t^{(i)})p(S_t^{(i)}|S_{1:t-1}^{(i)}, Z_{1:t}^{(i)})p(Z_t^{(i)}|Z_{1:t-1}^{(i)}) \tag{19}$$

where $S_{1:t-1}^{(i)}$ and $Z_{1:t-1}^{(i)}$ represent the value and cluster assignments in particle i, and n_i is the number of particles. The equations for each of these components are therefore identical to those given above restricted to the partition held by the particle under consideration. Once the probability of each valid assignment has been calculated, these probabilities are then used to stochastically sample new partitions including the latest observation to be held as the new particles for the next trial.

Similar processes can then be performed for both prediction and response selection, again restricting the considered permutations to those currently held in the particles; as such, the predictive distribution becomes:

$$p(S_{t+1}|X_{1:t}, F_{1:t}) \approx \frac{1}{n_i}\sum_i p(S_{t+1}^{(i)}|S_{1:t}^{(i)}, Z_{1:t+1}^{(i)})p(Z_{t+1}^{(i)}|Z_{1:t}^{(i)}) \tag{20}$$

replacing Eq. (16), while the response distribution becomes:

$$p(S_t|X_{1:t}, F_{1:t-1}) \approx \frac{1}{n_i}\sum_i p(X_t|S_t^{(i)})p(S_t^{(i)}|S_{1:t-1}^{(i)}, Z_{1:t}^{(i)})p(Z_t^{(i)}|Z_{1:t-1}^{(i)}) \tag{21}$$

replacing Eq. (18).

In addition to the particle filter, the model included a second approximation within the perceptual distribution of Eq. (4): to make computation more tractable, the sampled value X_t was replaced with the true value v_t, so assuming perceptual samples were perfectly accurate:

$$p(X_t|S_t) = logN(\ln(v_t), \sigma_l^2) \tag{22}$$

replacing Eq. (4). While this does remove some noise from the estimation system, this can be subsequently reinserted by sampling responses from the distribution given by Eq. (21) rather than simply taking the maximum, an approximation which has previously been found to be successful (e.g. Sanborn, Mansinghka, & Griffiths, 2013).

Finally, for the purposes of fitting the UEM to actual behaviour, the response distribution was further edited to include two additional elements: first, the distribution is raised to an exponent to allow the model to interpolate between probability matching and maximisation, and second, the response distribution is combined with a uniform background distribution to emulate potential noise in response selection:

$$p(R_t|X_{1:t}, F_{1:t-1}) = (1 - w_b)\frac{p(S_t|X_{1:t}, F_{1:t-1})^e}{\sum p(S_t|X_{1:t}, F_{1:t-1})^e} + w_b U(v_1, v_2) \tag{23}$$

where R_t is the potential response, e is the response exponent, w_b is the weight applied to the background distribution and v_1 and v_2 provide the range of values considered in the uniform distribution. Responses can then be drawn from this distribution using various methods, with the resulting feedback being used to update the representation using the above method. For the purposes of this study, however, no fixed sampling method is defined, with this distribution instead being used to provide the probability of a given participant response.

A.2. Model comparison

Full details of the model comparison procedure are provided here. As noted in the main text, the model comparison used a grid point search across model parameters to suggest best fits to the data. This was used in place of more traditional gradient descent functions due to potential issues with such methods for clustering models: the likelihood function of these models is often highly complex, leading gradient descent functions to become fixed at local maxima rather than the global maximum. The search ran across the four parameters shared by the two models: the coupling parameter c, response exponent e, feedback confidence c_f and background weight w_b. Considered values were: for c, 0.1 to 0.9 in steps of 0.1; for e, 0.1, 0.25, 0.5, 1, 1.5 and 2; for c_f, 0.1, 0.3, 0.5, 0.7 and 0.95 (capturing the stated accuracy in the 95% condition); and for w_b, 0.01, 0.1, 0.3, 0.5, 0.7 and 0.9.

In order to simplify the comparison, the prior parameters unique to the cUEM (μ_0, σ_0^2, β_0 and λ_0) were fixed across model fits. The values of these parameters were set according to the range of displayed dot counts following the format of Anderson (1991); however, in order to allow for the previously described emulation of categorical components by the Gaussian mixture prior, the prior variance and confidence values were edited to provide a narrower initial form. As such, the prior mean was set at the midpoint of the range (27.5), the prior variance was set at a twentieth of the range squared (0.2025), the confidence in the prior mean β_0 was set at one, and the prior variance confidence λ_0 was set at 0.01, determined through limited likelihood testing using manual adjustments of this parameter on a subset of the data. While these manipulations were limited, these were considered as full parameters for the purposes of calculating complexity penalties in subsequent measures. The dUEM was therefore defined as having four free parameters, while the cUEM had six.

Both models were then fit to each participant individually by providing the model with the observed dot counts in matching order for partitioning, calculating the response distribution (given in Eq. (23)) and taking the resulting probability of the participant's response for that trial. These trial probabilities were then converted into log likelihoods and summed for each participant at each grid point of each model. Maximum log likelihood figures for each model for each participant were then converted into AIC and BIC values, as used in the main text. Because of the stochasticity added to the models by the particle filter, each grid point was repeated

10 times to provide an average log likelihood to more reliably measure model fit, and the number of particles used within the filter was set to 20 across all grid points to encourage greater consistency between repetitions. These values were maximised whilst still making fitting computationally feasible within our available resources (using 25 new desktop computers each running four parallel sessions for one week), though this did not completely eliminate stochasticity in the simulations (as reported in the main text, average standard deviation in log likelihoods at the best fitting parameters was 199.86 for the dUEM and 213.36 for the cUEM).

As a check of the impact of this noise in likelihood estimates, additional simulations were subsequently performed at the best fitting parameters of each participant, providing 50 new repetitions at these points. Average log likelihood values across these additional simulations demonstrated high correlations with the initial estimates (dUEM: r = 0.942; cUEM: r = 0.923), but comparisons between the two models across these new simulations found mixed results: while the proportion of participants best fit by each model was consistent with the main analysis (dUEM: 25; cUEM: 74), the best fitting model differed for 33 of the 99 participants. As a result, we have been somewhat cautious in the conclusions drawn from this fitting process: while the collected data give reasonable confidence that the continuous model provides the better fit across all subjects, we cannot be similarly confident regarding the best fitting model for each individual participant. This seems to be due to the difficulty of fitting stochastic clustering processes such as those used in the present models to behaviour, as partitions can differ substantially even at identical parameters, leading to large differences in the estimated likelihood of observed data. The present results are therefore an approximation of the true best fits of the candidate models to participant data, though it is notable that the additional repetitions did find qualitatively similar results to the main analysis overall.

A.3. Additional modelling results

The following provides alternate model comparison results, beginning with the global fits assuming a common set of parameters across participants within each experiment, summarised in Table 4. This finds qualitatively identical results to the individual parameters reported in the main text, with the cUEM outperforming the dUEM in both experiments by a wide margin. As previously noted, however, fits are substantially better when using individual parameters in both tasks, making those findings more helpful in separating the models.

Table 4
Global modelling results from Experiments 1 and 2, where ΔMLL is the difference in maximum log likelihood for the best fitting model in that comparison assuming common parameters across participants in each experiment. Brackets give bootstrapped 95% CIs.

Experiment	Best Model	ΔMLL	ΔAIC	*w*(AIC)	ΔBIC	*w*(BIC)
Experiment 1	cUEM	1851 (2211, 1684)	−3699 (−3364, −4419)	1	−3682 (−3349, −4403)	1
Experiment 2	cUEM	2945 (3635, 2693)	−5904 (−5381, −7267)	1	−5868 (−5365, −7250)	1

Secondly, we here list the full results of the model comparisons using AIC values, as summarised in Table 3. For Experiment 1, as with the BIC scores above, aggregate AIC scores show the cUEM had a better fit to experimental data, with a similar proportion of participants being best fit by each model (dUEM: 5 [3–8 95% CI]; cUEM: 34 [31–36]). When separated by uncertainty condition, the cUEM held a better fit in both the 75% and 95% groups, suggesting behaviour was best described using a Gaussian mixture prior even where feedback was more reliable; this is further displayed in the ratios of participants best fit by each model, with the cUEM accounting for 16 (14–19) of the 19 participants in the 70% condition and 18 (16–19) of the 20 participants in the 95% condition. As with the BIC measures, this difference in ratio between conditions was non-significant, $\chi^2(1) < 0.01$, p = .951.

For Experiment 2, aggregate AIC scores also found the cUEM held a better fit to data, accounting for 52 (48–54) of the 60 participants. When divided by uncertainty condition, the cUEM again held a better fit in both the 75% and 95% groups, accounting for 27 (24–28) of the 30 participants in the 70% condition and 25 (23–27) of the 30 participants in the 95% condition. Again, this ratio did not significantly differ between groups, $\chi^2(1) = 0.14$, p = .704.

A.4. Parameter values

We here provide more detailed discussion of the results of the model fitting based on the best-fitting parameters suggested by the model comparison, offering further insight on the behaviour inferred by the candidate models. Fig. 8 shows the distribution of values taken from the best fits of both models to participant data. These distributions show reasonably similar patterns between the models, though there are some notable distinctions: first, the coupling parameter was generally lower for the discrete model, indicating a tendency towards a larger number of clusters; this is understandable given that clusters in the discrete model are less varied, meaning a wider set of clusters may be required to adequately represent the target distribution. Second, the response exponent is reasonably similar between models, but tended to be slightly higher for the continuous model, suggesting a tendency towards probability matching in response selection for that model. Third, feedback confidence is reasonably high in both models, indicating that participants did make use of feedback information despite its potential inaccuracy, though it is notable that a number of participants used lower confidence values than the accuracy level stated in the experimental instructions. Finally, weight on the uniform background distribution was low for both models, suggesting little noise in responses.

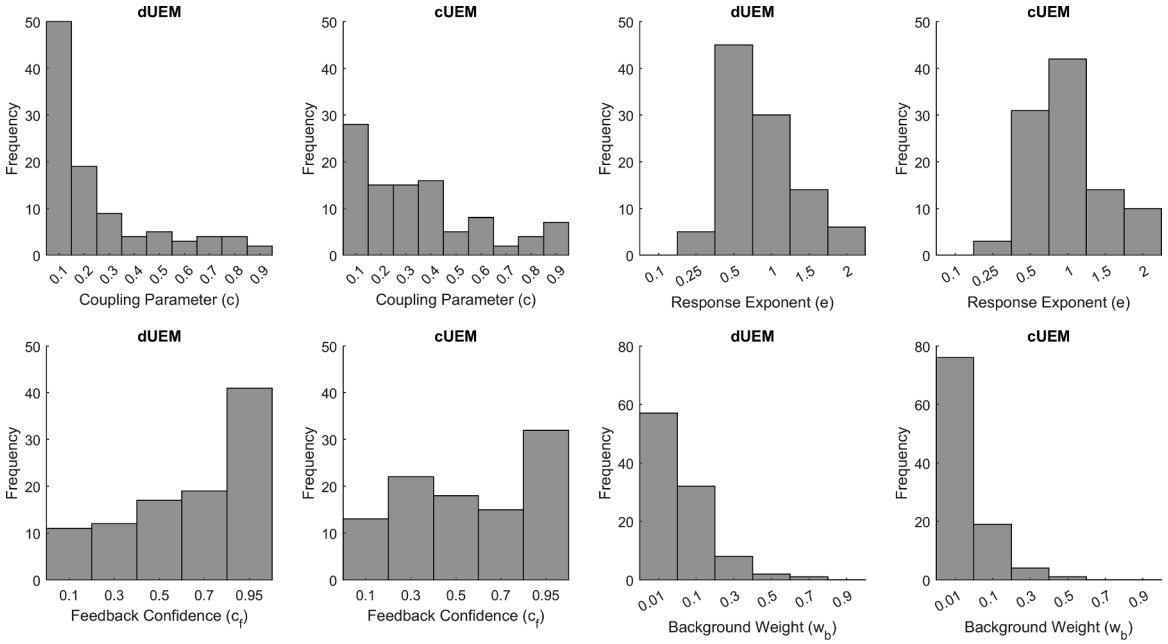

Fig. 8. Histograms of the best-fitting parameter values from the considered figures given to both models collected across all participants from the two experiments.

A.5. Model lesioning

To test the actual impact of the use of these prior distributions on the accuracy of subsequent estimation, the continuous and discrete models described above were compared with a lesioned version of the UEM removing either prior, labelled the lUEM. This meant that responses were based solely on perceptual data, as defined by Eq. (22), though this distribution was again modified by the response exponent and background distribution as in Eq. (23):

$$p(R_t|X_t) = (1 - w_b)\frac{logN(log(v_t), \sigma_l^2)^e}{\sum logN(log(v_t), \sigma_l^2)^e} + w_b U(v_1, v_2) \tag{24}$$

The dUEM, cUEM and lUEM were run at the best fitting parameters found for each model for each participant in the above model comparison and used to calculate an estimate of accuracy by taking the average probability of the model giving the true displayed value as a response across estimate trials. The predicted accuracy of the lUEM was significantly lower than both the dUEM ($t(99) = 16.9$, $p < .001$) and cUEM ($t(99) = 15.9$, $p < .001$), suggesting the use of either the discrete or continuous prior distributions benefits estimation performance. This is understandable as the lesioned model by definition has no knowledge of the underlying prevalence of values, and so is unable to capture complex distributions such as those used in the present experiments.

A.6. Correspondence between the discrete mixture and the Dirichlet distribution

We here provide a comparison between the definition of the discrete mixture prior used in this study and a distribution commonly used as a prior in for discrete formats, the Dirichlet distribution. These two distributions can in fact be shown to be equivalent, as demonstrated below. As such, any conclusions regarding the discrete model made in this study can also be applied to the Dirichlet prior, most prominently that both appear to be outperformed by the use of a continuous numerical representation.

Using the definition of the discrete mixture prior given in Eq. (25), we first substitute in $p(S_t|S_{1:t-1}, Z_{1:t})$ and $p(Z_t|Z_{1:t-1})$ to produce Eq. (26), using a form of the latter term that first gives the prior probability of Z_t being new combined with a uniform probability over all K possible alternatives. The second part of $p(Z_t|Z_{1:t-1})$ sums over all of the old cluster Z_t, weighting each with 0 or 1 (e.g., n_s/n_z) depending on whether that cluster includes dot counts equal to the current observation S_t. In Eq. (27), the weights and priors for each old cluster Z_t have been summed over, and $n_{s'}$ is simply the number of past dot counts equal to S_t regardless of assignment to clusters. As the probabilities of S_t now no longer depend on past assignments $Z_{1:t-1}$, these assignments are dropped (as $\sum_{Z_{1:t-1}} p(Z_{1:t-1}|S_{1:t-1}) = 1$) in Eq. (28). Finally, we define $a_0 = (1 - c)/c$ and reorder the terms to produce Eq. (29). Eq. (29) shows that $p(S_t|S_{1:t-1})$ is the conditional probability of S_t given the past dot counts and a symmetric Dirichlet prior with parameters a_0/K over all of the K possible responses.

$$p(S_t|S_{1:t-1}) = \sum_{Z_{1:t}} p(S_t|S_{1:t-1}, Z_{1:t})p(Z_t|Z_{1:t-1})p(Z_{1:t-1}|S_{1:t-1}) \tag{25}$$

$$= \sum_{Z_{1:t-1}} \left(\left(\frac{(1-c)}{(1-c)+cn} \right) \left(\frac{1}{K} \right) + \sum_{Z_t \in Z_{1:t-1}} \left(\frac{n_s}{n_z} \right) \left(\frac{cn_z}{(1-c)+cn} \right) \right) p(Z_{1:t-1}|S_{1:t-1}) \tag{26}$$

$$= \sum_{Z_{1:t-1}} \left(\frac{(1-c)}{K((1-c)+cn)} + \frac{cn_{s'}}{(1-c)+cn} \right) p(Z_{1:t-1}|S_{1:t-1}) \tag{27}$$

$$= \frac{(1-c)+Kcn_{s'}}{K((1-c)+cn)} \tag{28}$$

$$= \frac{n_{s'}+\alpha_0/K}{n+\alpha_0} \tag{29}$$

References

Acerbi, L., Vijayakumar, S., & Wolpert, D. M. (2014). On the origins of suboptimality in human probabilistic inference. *PLOS Computational Biology, 10*, e1003661. https://doi.org/10.1371/journal.pcbi.1003661.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723. https://doi.org/10.1109/TAC.1974.1100705.

Aldous, D. (1985). Exchangeability and related topics. *École d'été de probabilités de Saint-Flour, XIII-1983* (pp. 1–198). Berlin: Springer.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98*, 409–429. https://doi.org/10.1037/0033-295X.98.3.409.

Buchsbaum, D., Griffiths, T. L., Plunkett, D., Gopnik, A., & Baldwin, D. (2015). Inferring action structure and causal relationships in continuous sequences of human action. *Cognitive Psychology, 76*, 30–77. https://doi.org/10.1016/j.cogpsych.2014.10.001.

Chalk, M., Seitz, A. R., & Series, P. (2010). Rapidly learned stimulus expectations alter perception of motion. *Journal of Vision, 10*, 2. https://doi.org/10.1167/10.8.2.

Dehaene, S. (2011). *The number sense: How the mind creates mathematics.* New York: Oxford University Press.

Dehaene, S., & Marques, J. F. (2002). Cognitive euroscience: Scalar variability in price estimation and the cognitive consequences of switching to the euro. *Quarterly Journal of Experimental Psychology, 55*, 705–731. https://doi.org/10.1080/02724980244000044.

Ditz, H. M., & Nieder, A. (2016). Numerosity representations in crows obey the Weber-Fechner law. Proceedings of the Royal Society B - Biological Sciences 283, 20160083. doi:10.1098/rspb.2016.0083.

Flombaum, J. I., Junge, J. A., & Hauser, M. D. (2005). Rhesus monkeys (Macaca mulatta) spontaneously compute addition operations over large numbers. *Cognition, 97*, 315–325. https://doi.org/10.1016/j.cognition.2004.09.004.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis.* London: CRC Press.

Gershman, S. J., & Niv, Y. (2013). Perceptual estimation obeys Occam's Razor. *Frontiers in Psychology, 4*, 623. https://doi.org/10.3389/fpsyg.2013.00623.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition, 112*, 21–54. https://doi.org/10.1016/j.cognition.2009.03.008.

Griffiths, T. L., Sanborn, A. N., Canini, K. R., Navarro, D. J., & Tenenbaum, J. B. (2011). Nonparametric Bayesian models of categorization. In E. M. Pothos, & A. J. Wills (Eds.). *Formal approaches in categorization* (pp. 173–198). Cambridge, UK: Cambridge University Press.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science, 21*, 263–268. https://doi.org/10.1177/0963721412447619.

Hayes, B. K., Banner, S., & Navarro, D. J. (2017). Sampling frames, Bayesian inference and inductive reasoning. In G. Gunzelmann, A. Howes, T. Tenbrink, & E.J. Davelaar (Eds.), Proceedings of the 39th annual conference of the cognitive science society. Austin, TX: Cognitive Science Society.

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

Kording, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature, 427*, 244–247. https://doi.org/10.1038/nature02169.

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

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* New York: Freeman.

McCrink, K., & Wynn, K. (2004). Large-number addition and subtraction by 9-month-old infants. *Psychological Science, 15*, 776–781. https://doi.org/10.1111/j.0956-7976.2004.00755.x.

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature, 215*, 1519–1520. https://doi.org/10.1038/2151519a0.

Pitman, J. (2002). Combinatorial stochastic processes. (Notes for Saint Flour Summer School).

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies – Revisited. *Neuroimage, 84*, 971–985. https://doi.org/10.1016/j.neuroimage.2013.08.065.

Rosseel, Y. (2002). Mixture models of categorization. *Journal of Mathematical Psychology, 46*, 178–210. https://doi.org/10.1006/jmps.2001.1379.

Sanborn, A. N., & Beierholm, U. R. (2016). Fast and accurate learning when making discrete numerical estimates. *PLoS Computational Biology, 12*, e1004859. https://doi.org/10.1371/journal.pcbi.1004859.

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Science, 20*, 883–893. https://doi.org/10.1016/j.tics.2016.10.003.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review, 117*, 1144–1167. https://doi.org/10.1037/a0020511.

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological Review, 120*, 411–437. https://doi.org/10.1037/a0031912.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464. https://doi.org/10.1214/aos/1176344136.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology, 71*, 55–89. https://doi.org/10.1016/j.cogpsych.2013.12.004.

Spelke, E. S., & Tsivkin, S. (2001). Language and number: A bilingual training study. *Cognition, 78*, 45–88. https://doi.org/10.1016/S0010-0277(00)00108-6.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *Neuroimage, 46*, 1004–1017. https://doi.org/10.1016/j.neuroimage.2009.03.025.

Sun, J., Li, J., & Zhang, H. (2019). Human representation of multimodal distributions as clusters of samples. *PLoS Computational Biology, 15*, e1007047. https://doi.org/10.1371/journal.pcbi.1007047.

Tauber, S., Navarro, D. J., Perfors, A., & Steyvers, M. (2017). Bayesian models of cognition revisited: Setting optimality aside and letting data drive psychological theory. *Psychological Review, 124*, 410–441. https://doi.org/10.1037/rev0000052.

Vanpaemel, W., & Storms, G. (2008). In search of abstraction: The varying abstraction model of categorization. *Psychonomic Bulletin and Review, 15*, 732–749. https://doi.org/10.3758/PBR.15.4.732.

Wagenmakers, E., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin and Review, 11*, 192–196. https://doi.org/10.3758/BF03206482.