Supporting Information

Samuel J. Cheyette, Shengyi Wu, Steven T. Piantadosi

Model derivation

Using Lagrange multipliers to find the optimal Q

We find a form of $Q(s' \mid s)$ chosen to minimize the expected loss between an input s and its representation s',

$$\mathbb{E}\left[\mathcal{L}(s, s')\right] = \sum_{s \in R} P(s) \sum_{s' \in R} Q(s' \mid s) \cdot \mathcal{L}(s, s'). \tag{1}$$

Here $P(s)$ is the prior probability of encountering the scene s. Assuming that B is the maximum allowable information, we optimize (1) subject to a bound on the KL-divergence between Q and P,

$$D_{KL}\left[Q(\cdot \mid s) \parallel P(\cdot)\right] = \sum_{s' \in R} Q(s' \mid s) \cdot \log \frac{Q(s' \mid s)}{P(s')} \leq B \quad \forall s \in R. \tag{2}$$

where R is the set of all possible scenes.

Since Q is a distribution, we also have a constraint that $\sum_{s' \in R} Q(s' \mid s) = 1$ for all s.

To apply the method of Lagrange multipliers, we encode the objective function and constraints into a single equation,

$$\begin{aligned}\mathcal{F}[Q(s' \mid s)] = &\sum_{s \in R} P(s) \sum_{s' \in R} Q(s' \mid s) \cdot L(s, s') \\ &+ \sum_{s \in R} \lambda_s \cdot \left(B - \sum_{s' \in R} Q(s' \mid s) \log \frac{Q(s' \mid s)}{P(s')} \right) \\ &+ \sum_{s \in R} \gamma_s \cdot \left(1 - \sum_{s' \in R} Q(s' \mid s) \right).\end{aligned}$$

We then solve for the of the zeroes of the derivative of $\mathcal{F}$ with respect to $Q(s' \mid s)$ (i.e. treating "$Q(s' \mid s)$" as a separate variable for each s and s'). These zeros occur when

$$P(s) \cdot L(s, s') + \lambda_s \cdot \left(1 + \log \frac{Q(s' \mid s)}{P(s')} \right) + \gamma_s = 0 \tag{3}$$

or

$$Q(s' \mid s) \propto P(s') \cdot \exp\left(-\frac{P(s)}{\lambda_s} \cdot L(s, s')\right). \tag{4}$$

Here, λ_s is chosen to satisfy the bound in (2).

Finding λ_s using numerical approximation

We solve for λ_s using numerical methods. Specifically, given an information bound, we used gradient descent to find λ_s that allows the maximum $D_{KL}\left[Q(\cdot|s) \parallel P(\cdot)\right]$ that satisfies the constraint. This optimizer was run for 5,000 steps for each λ_s, which is sufficient to find KL-divergences within a millionth of a bit of the bound.

One complication is that the representational space in our experiments was very large — there are 49 grid cells so there are 2^{49} possible grid states ($\approx 10^{15}$). Memory and runtime constraints therefore make it impossible to represent the prior and posterior of each possible grid state independently. Luckily, for every scene s, there are many representations that are "equivalent" in that they have equal prior probabilities and losses. For a given representation s', we define the loss as a function of the number (or proportion) of false positives and false negatives between s and s'. To get the number of false negatives $f_n(s' \mid s)$ and false positives $f_p(s' \mid s)$, we can write,

$$f_n(s' \mid s) = \sum_i \sum_j s_{ij} \cdot (1 - s'_{ij}), \tag{5}$$

and

$$f_p(s' \mid s) = \sum_i \sum_j (1 - s_{ij}) \cdot s'_{ij}, \tag{6}$$

where i and j are the rows and columns of the grid.

We can count the number of representations that have $f_n(\cdot \mid s) = r_n$ and $f_p(\cdot \mid s) = r_p$. This is the product of all the ways to make $n - r_p$ true positives in given that s in n on cells and $k - r_n$ true negatives in $M - n$ off cells, where n is again the cardinality of the scene s, k is the cardinality of the representation s', and M is the total number of grid cells. So we therefore can write the total number of equivalent states S as,

$$S = \binom{n}{n - r_p}\binom{k - r_n}{M - n}. \tag{7}$$

In this way, we can collapse the representational space into only individual instances of each equivalence class and when calculating the KL-divergence multiply each term by S.

Change-localization task

We assume that subjects choose in the change-localization task proportionally to their belief that a cell has changed. In disappearing trials, subjects are only allowed to respond with a zero cells, and in this case the probability that the cell changed is the belief that the cell was initially 1. This means that the probability of responding ij out of only the other zeros is,

$$P(\text{choose } ij) \propto \sum_{s' \in R} Q(s' \mid s) \cdot \mathbf{1}_{s'_{ij}=1}. \tag{8}$$

To compute the probability that subjects answer accurately, $P(\text{choose } ij)$ is computed for the correct disappearing cell relative to all of the other zero cells in the final display. Appearing trials are defined analogously.

Numerical estimation task

To compute the probability the model believes that the scene contained k' objects, we can sum across the model's posterior for all scenes containing k' objects. More formally,

$$p(k = k' \mid s) = \sum_{s' \in R} Q(s' \mid s) \cdot \mathbf{1}_{|s'|=k'} \tag{9}$$

where $|s'|$ represents the cardinality of representation s' (i.e. the number of objects in s'), and $\mathbf{1}_{x=y}$ is 1 when $x = y$ and 0 otherwise.

Model fitting

For both experiments, we used a Markov Chain Monte Carlo (MCMC) algorithm to fit four parameters to the data: a) power law functions for how the information capacity changes over time, of the form $a \cdot t^k$, with a and k as free parameters and t representing time in seconds; b) the loss function parameter α, which weights the cost of false negatives and false positives; and c) a guessing parameter p_g which captured the rate of choosing randomly. Because α and p_g represented probabilities and thus were constrained to be between 0 and 1, we parameterized these through transformations $\alpha' = logit(\alpha)$ and $p'_g = logit(p_g)$. We fit these parameters in a hierarchical Bayesian network, with partial pooling of parameter estimates across participants. We used uninformative group-level priors for the means of each parameter, which we believed would not exert a strong influence in any case given the large amount of data collected. We drew group-level standard deviations from $HalfNormal(\sigma = 10)$. Subjects' parameters were drawn from the distributions,

$$a_s \sim Normal(\mu_{a,g}, \sigma_{a,g}), \tag{10}$$

$$k_s \sim Normal(\mu_{k,g}, \sigma_{k,g}), \tag{11}$$

$$\alpha'_s \sim Normal(\mu_{\alpha',g}, \sigma_{\alpha',g}), \tag{12}$$

$$p'_{g_s} \sim Normal(\mu_{p'_g,g}, \sigma_{p'_g,g}), \tag{13}$$

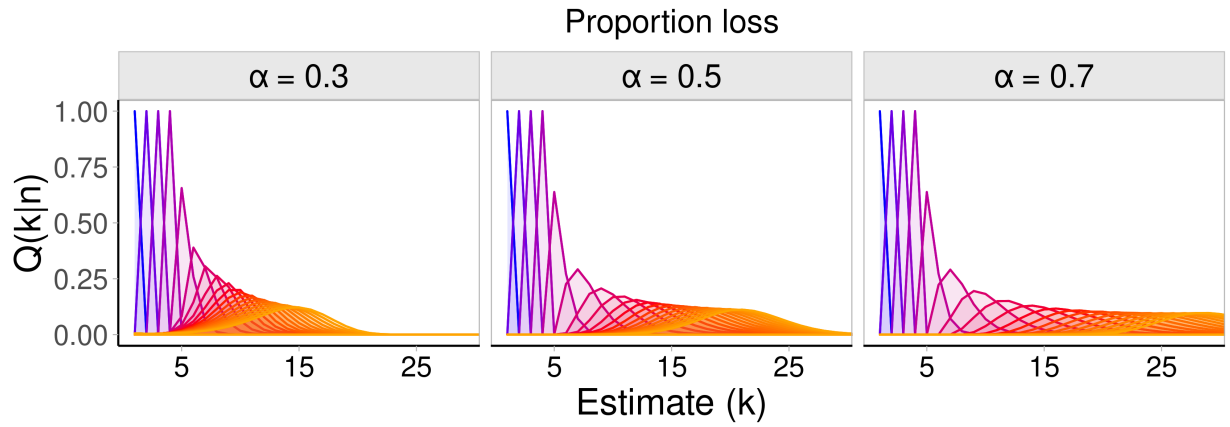
where group-level parameters are denoted $\mu_{.,g}$ and $\sigma_{.,g}$ and subject-level parameters are denoted with subscript s.

We used the Metroplolis-Hastings algorithm to jointly fit the posterior distributions of each group-level and subject-level parameter. Because there is a high runtime cost to compute the model's posterior distribution, we rounded the information bounds given by samples of a and k to the nearest 0.1, and each α to the nearest 0.01, and cached the results. This can only have a negative impact on the fit of the model and so it could not impact (e.g.) model comparisons in our model's favor. We ran two chains of Metropolis-Hastings for 50,000 steps, with 10,000 steps of burn-in, storing every 10th value to avoid auto-correlation of samples. We checked for convergence of the chains using the Gelman-Rubin statistic *(63)*, and found in both tasks that $\hat{r} < 1.05$ for all group-level parameters and $\hat{r} < 1.1$ for all subject-level parameters, indicating that the chains converged.

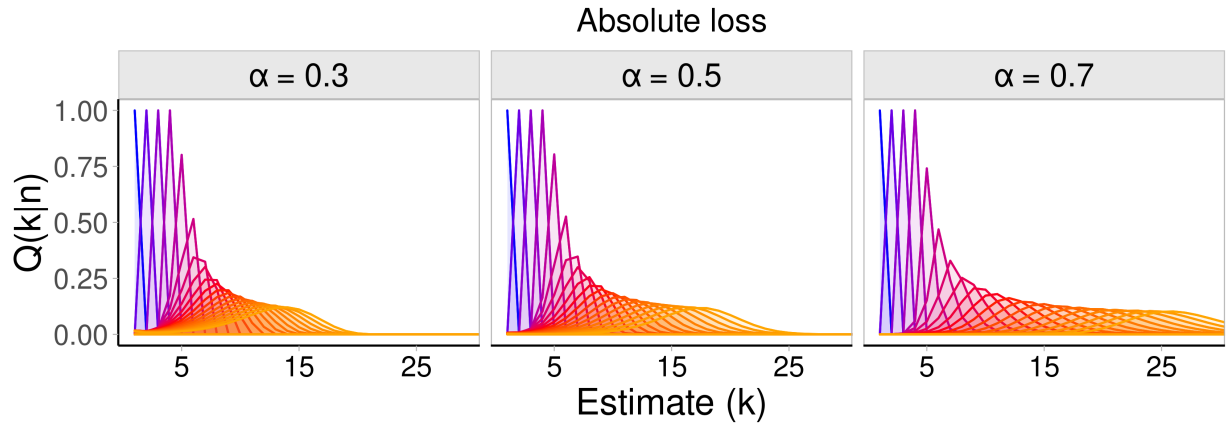
Alternative loss functions

In the main text, we used a loss function that combined a weighted proportion of false negatives and false positives relative to the number of locations with objects and locations without objects respectively. We had pre-registered this choice, however, it is not the only plausible loss function. One alternative choice would be the total number of places the representation s' differs from the scene a; another would be a possibly weighted combination of the *number* rather than *proportion* of false negatives and false positives. Here we show that while the choice of loss function somewhat influences the form of the resulting psychophysics, the outcomes are qualitatively very similar and preserve the core properties of the model in the paper.

To consider these loss functions, here we will define terms slightly differently than in the main text. For a given scene s and representation s' we will define a function for the number of false negatives $f_n(s' \mid s)$ and false positives $f_p(s' \mid s)$. We can write,



(a)



(b)

Fig. S1: These two plots illustrate the number psychophysics produced by various formulations of the loss function. Each line shows the estimates produced for a different number $n = 1...20$. We assume an information bound of 25 bits. (a) These panels illustrate the psychophysics produced by different parameterizations of the loss function assumed in the main text, weighting the proportion of false negatives out of true positives by alpha and weighting the proportion of false positives out of the true negatives by one minus alpha. Each panel shows a different possible weighting, with $\alpha = 0.3$, $\alpha = 0.5$, and $\alpha = 0.7$. (b) These panels illustrate the psychophysics assuming a loss function that is an analogous weighted combination of the *number* rather than *proportion* of false negatives and false positives.

$$f_n(s' \mid s) = \sum_i \sum_j s_{ij} \cdot (1 - s'_{ij}), \tag{14}$$

and

$$f_p(s' \mid s) = \sum_i \sum_j (1 - s_{ij}) \cdot s'_{ij}. \tag{15}$$

Then we can write the loss function assumed in the paper (using proportions) as,

$$\mathcal{L}_{proportion}(s, s') = \alpha \cdot \frac{f_n(s' \mid s)}{k} + (1 - \alpha) \cdot \frac{f_p(s' \mid s)}{n - k}. \tag{16}$$

The loss function that is a weighted combination of the number, rather than proportion, can be written as,

$$\mathcal{L}_{absolute}(s, s') = \alpha \cdot f_n(s' \mid s) + (1 - \alpha) \cdot f_p(s' \mid s). \tag{17}$$

Figure S1 shows predicted number psychophysics using both loss functions under different values of α, with Figure S1a showing the proportional loss function used in the main text and Figure 1b showing the absolute numeric loss function. At $\alpha = 0.5$ (middle panels), the weighting of both false negatives and false positives (either by proportion or absolute value) is equal; false negatives are under-weighted on the left panels and over-weighted on the right panels. Comparing the loss functions at each value of α, the psychophysics look very similar, particularly for low values of α. At higher values of α, the proportional loss function over-weights false negatives more strongly than the numeric counterparts for large numbers, and so ends up over-estimating.

Parameter recovery

In order to determine that the model parameters are recoverable, we simulated data from 200 "participants" with different values of the information-rate parameters a and k and different loss function parameters α. The a parameters were randomly sampled from Uniform$(10, 90)$; the k parameters were randomly sampled from Uniform$(0.01, 0.99)$; and the α parameters were randomly sampled from Uniform$(0.1, 0.9)$. The guessing rate was fixed to 0.01. Data was then generated by sampling estimates and localization guesses for each trial from Experiment 3 and maximum likelihood fitting was performed to recover the parameters separately for the estimation data and localization data.

Each parameter could be recovered with relatively high fidelity. For the estimation data, there were correlations of $r = 0.92$ for a and inferred $\hat{a}$; $r = 0.87$ for k and inferred $\hat{k}$; and $r = 0.96$ for α and inferred $\hat{\alpha}$. For the localization data, there were correlations of $r = 0.91$ for a and inferred $\hat{a}$; $r = 0.76$ for k and inferred $\hat{k}$; and $r = 0.87$ for α and inferred $\hat{\alpha}$. Figure S2 shows the true parameters versus recovered parameters for the estimation task (a-c) and localization task (d-f). These plots show that recovered parameters largely lie on the $y = x$ line, meaning there is not significant bias in parameter estimates.

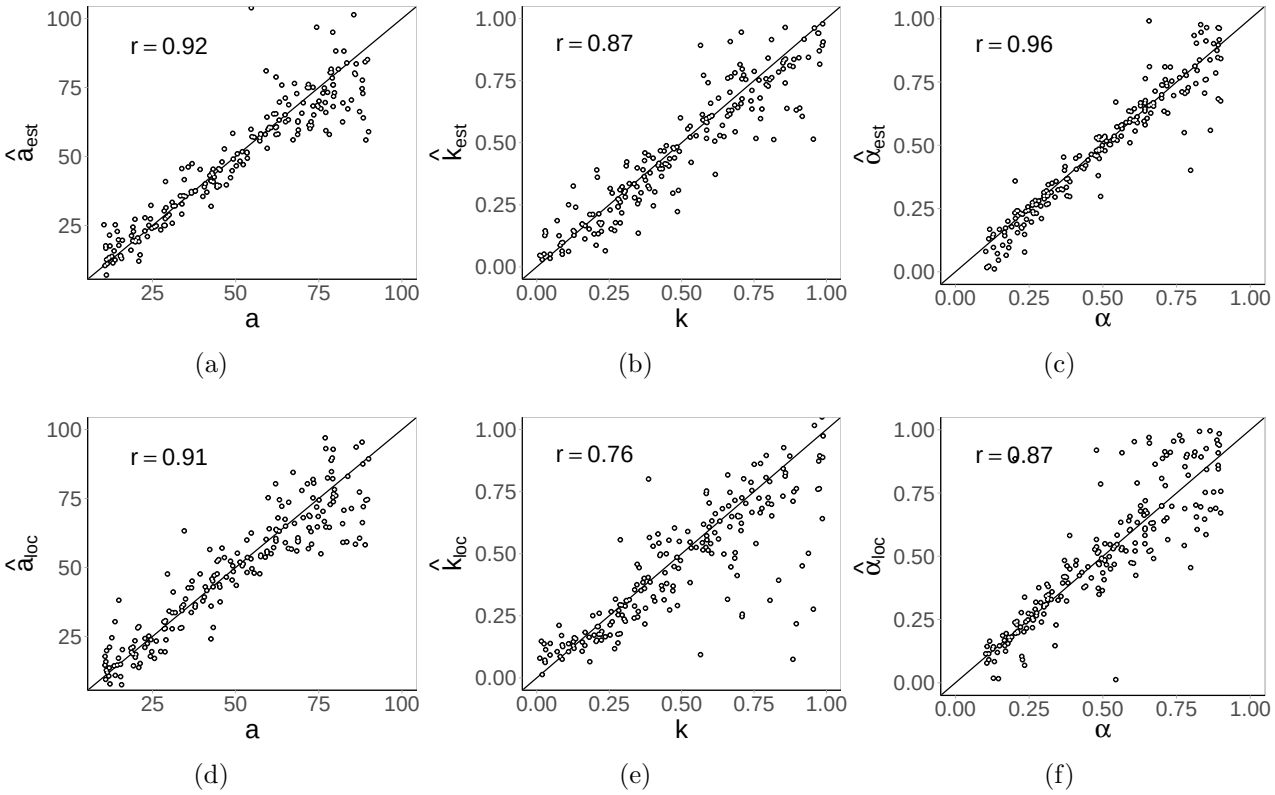


Fig. S2: Parameter recovery for simulated data. Panels a-c show maximum likelihood fits from simulated participants who each performed 45 trials of the estimation task, with (a) showing inferred a relative to inferred $\hat{a}$, (b) showing inferred k relative to $\hat{k}$; and (c) showing α relative to inferred $\hat{\alpha}$. Panels $d - f$ show analogous comparisons of inferred parameters relative to simulated parameters for the localization task.

Comparison of model fit with and without α

Model fitting recovered relatively low values of α (0.31-0.35) in all experiments (see Main Text). When we performed fits that constrained α to 0.5, the model did not fit as well overall or for most participants. In Experiment 1 (change-localization), the sum ΔAIC was 223 in favor of the model with freely varying α; 62/100 participants had lower AIC when α was fit. In Experiment 2 (estimation), the sum ΔAIC was 1,410; and 99/100 participants had lower AIC values when α was allowed to freely vary. Similarly, in Experiment 3, the sum ΔAIC was 855 in favor of the freely-varying model; all participants had lower AIC values when α was allowed to freely vary. In Experiment 3, 83 of the 100 participants had lower ΔAIC in the change-localization task and 100/100 had lower ΔAIC in the estimation task.

The effect of the prior

In the Cheyette & Piantadosi (2020) model, the decreasing prior over numerosities plays the central role in determining the noise and bias of estimates as a function of magnitude. That model would therefore predict that if a large number, say 75, happened to be high in the prior, people should be able to accurately represent sets of 75 items. But this seems perceptually implausible — could people really represent 75 items with higher fidelity than 2 items? One possible way of understanding the intuition that large groups of objects are intrinsically more difficult to represent precisely than smaller groups is that there is a lot more *spatial* information to represent about large groups.

If we take the simple method used in this paper of dividing the world up into a grid with M possible locations, then there are $\binom{M}{n}$ ways to represent n objects in space. There are M places to put a single object, meaning it takes only $\log M$ bits to represent scenes when $n = 1$. However, there are many more ways to place n items when n grows larger (as it approaches its zenith at $\frac{M}{2}$). Using Stirling's approximation of the Binomial, it takes about $\log \frac{4^n}{\sqrt{\pi n}}$ bits to represent $\frac{M}{2}$ objects. To put this in perspective, if $M = 50$, it would take $\log 50 \approx 5.6$ bits to represent $n = 1$ object's location but about $\log \frac{4^{25}}{\sqrt{25\pi}} \approx 47$ bits to represent the location of $n = 25$ objects.

Unlike Cheyette & Piantadosi (2020), the model we present in this paper accords with the intuition that more numerous sets are intrinsically more difficult to process perceptually. Even if there were a uniform prior over numerosities, small numerosities would be represented with significantly higher fidelity. In fact, the shape of the prior has much less of an impact on either mean estimates or the standard deviation of estimates relative to the loss function. We demonstrate this property in Figures S3-S5.

Suppose the prior on a scene s with n objects is given by the function $P(s \mid |s| = n) \propto 1/\left(n^\beta \cdot \binom{M}{n}\right)$, where β is a free parameter controlling the numerical bias. So $\beta = 2$ here is the naturalistic need frequency of number used in the paper ($P(n) \propto 1/n^2$) and $\beta = 0$ corresponds to a uniform prior over numerosities. Figures S3-S5 give the model's predictions for mean estimates and standard deviations under these two distributions ($\beta = 0$ and $\beta = 2$), at different values of the loss function parameter α (controlling how much the model cares about false positives versus false negatives).

Figure S3 demonstrates that the bias in the model's mean estimates is affected much more strongly by α than by β — i.e., the loss function, rather than the prior, mostly determines the patterns of under- or over-estimation. Figure S4 shows, analogously, the model's predictions for the coefficient of variation (CV) as a function of numerosity ($CV = \frac{\sigma}{\mu}$). Crucially, Fig-

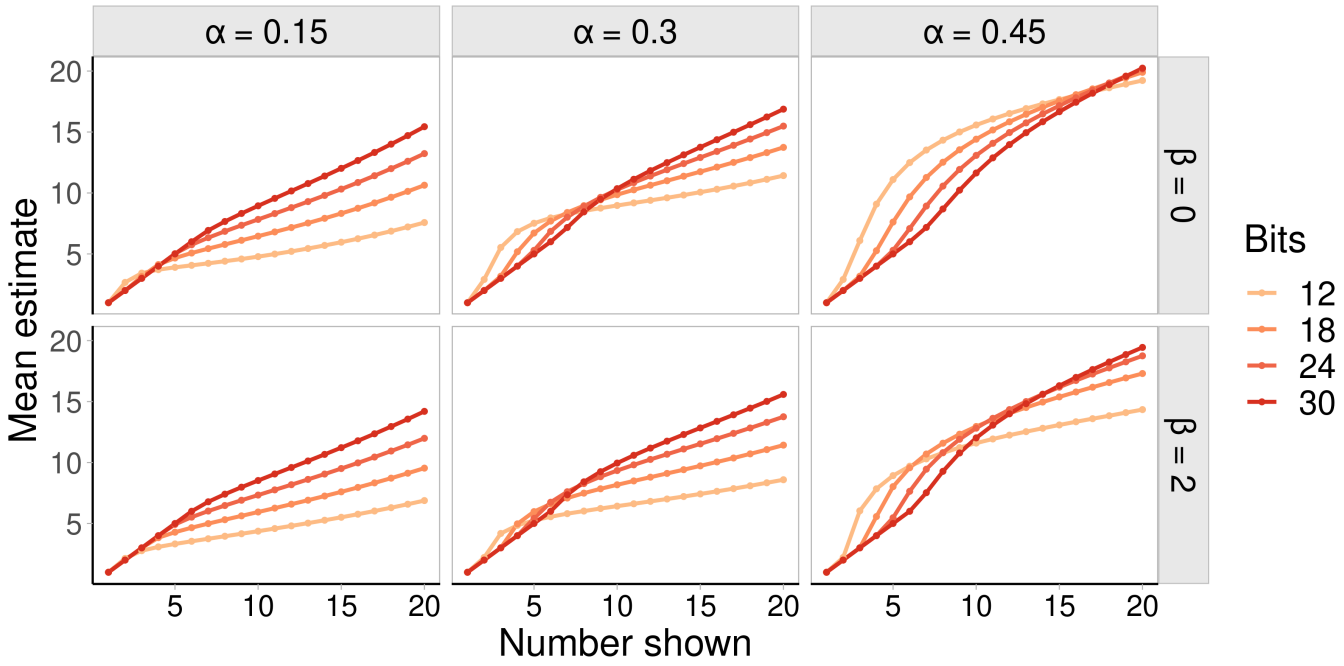


Fig. S3: Predicted mean estimates as a function of the number of objects shown (x-axis) and the information bound (color). The columns give predictions under different loss function parameters (α) and the rows show predictions for a uniform prior distribution ($\beta = 0$) and naturalistic need frequency ($\beta = 2$) used in the main text.

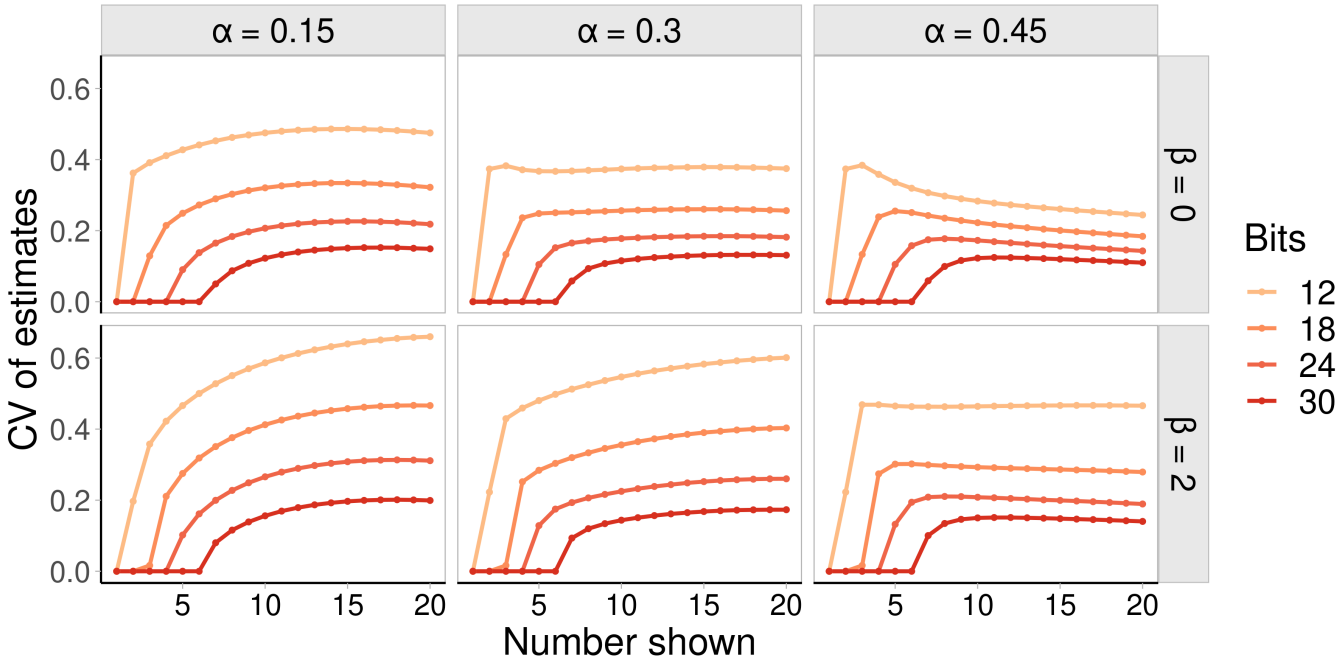


Fig. S4: Predicted coefficient of variation ($CV = \frac{\sigma}{\mu}$) as a function of the number of objects shown (x-axis) and the information bound (color). The columns give predictions under different loss function parameters (α) and the rows show predictions for a uniform prior distribution ($\beta = 0$) and naturalistic need frequency ($\beta = 2$) used in the main text.

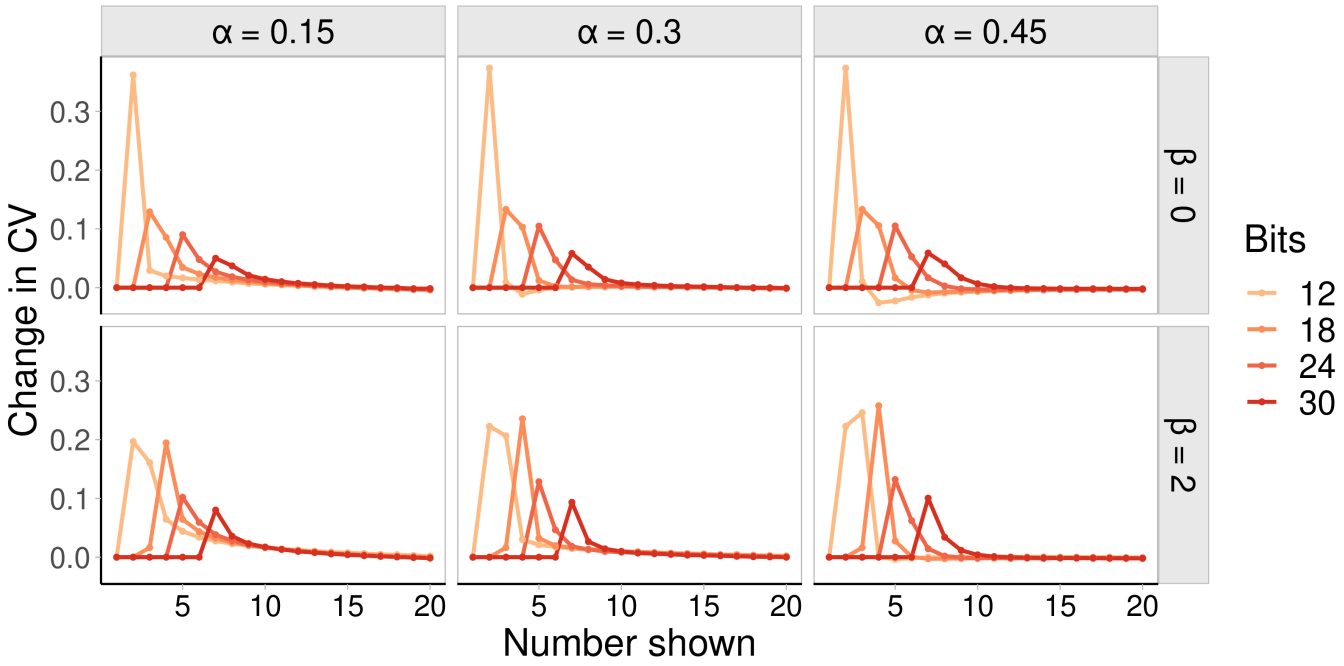


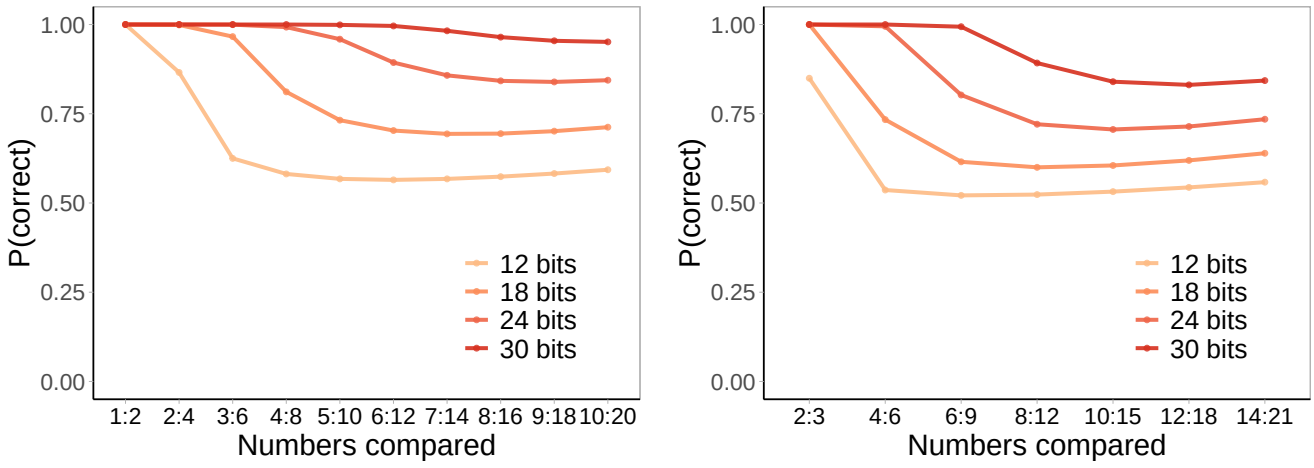
Fig. S5: Predicted change in the coefficient of variation ($\delta CV = CV_n - CV_{n-1}$) as a function of the number of objects shown (x-axis) and the information bound (color). The columns give predictions under different loss function parameters (α) and the rows show predictions for a uniform prior distribution ($\beta = 0$) and naturalistic need frequency ($\beta = 2$) used in the main text.

ure S4 illustrates that even with a uniform prior ($\beta = 0$), the model precisely represents small numerosities but not larger ones. In fact, the point of transition from subitizing to estimation is essentially entirely determined by the information bound, with α and β only having any significant influence on the standard deviation of estimates beyond the subitizing range.

Finally, Figure S5 demonstrates that the change in CV converges to 0 for larger numerosities, across different choices of the prior and loss function. This indicates that the model recovers Weber's law in estimation — which predicts a constant CV across numerosities above the subtizing range — without requiring fine-tuning of any parameters. A further demonstration that the model recovers Weber's law in estimation is given in the section below.

Weber's law

In addition to an estimation task, the model can be extended to a numerical discrimination task. For two numbers n_1 and n_2, we make model predictions for n_1 and n_2 independently and subsequently compute the probability that the model believes that n_2 was greater in magnitude



(a) (b)

Fig. S6: Model predictions for numerical discrimination on (a) 1:2 ratios and (b) 2:3 ratios. The model was parameterized with $\alpha = 1/3$ and the prior used in the main text.

than n_1,

$$P(n_2 > n_1) = \sum_{k=1}^{M-1} \sum_{j=k+1}^{M} P(k \mid n_1) \cdot P(j \mid n_2). \tag{18}$$

Figure S6 shows model predictions for discrimination performance on 1:2 ratios for numerosities 1:2 through 10:20 (a) and 2:3 ratios for numerosities 2:3 through 14:21 (b) across information capacity bounds. Weber's law implies that performance should be constant across ratios, which is true for the model somewhat beyond the subitizing range.

The relationship between subitizing and estimation

As discussed in the main text, previous work has shown that the relationship between subitizing and estimation is not straightforward. For instance, subitizing seems to be more greatly affected by attentional load than estimation (64); other studies have found little or no correlation between one's subitizing range and their estimation acuity (e.g. 2). One possible explanation afforded by the model is that small changes in capacity can lead to sharp changes in the subitizing range. Conversely, changes in capacity can lead to no changes in the subitizing range whatsoever. This could lead to puzzling results — subitizing and estimation will sometimes seem related but sometimes not. But, as we show, the model actually predicts that large changes in capacity are necessary for the relationship to become apparent.

We modeled the relationship between estimation acuity and subitizing range with the range of numerosities (1-8) tested in the studies cited above (2, 64). The subitizing range was calcu-

lated as the largest number with $\epsilon < 0.001$ squared estimation error; and the estimation acuity was calculated as the average coefficient of variation of numerosities beyond the subitizing range. Figure S7 shows the results of this simulation, with the subitizing range on the x-axis, estimation acuity acuity on the y-axis, and each point representing the model's prediction at a given information capacity. There are sudden changes in the subitizing range as the information capacity increases; conversely, there are small, less dramatic effects on estimation acuity.

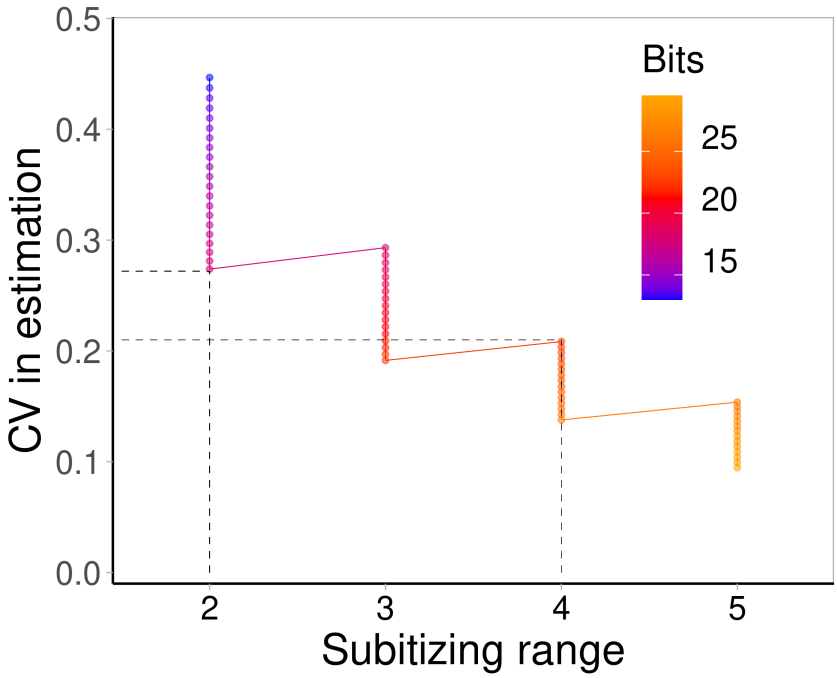


Fig. S7: The relationship between subitizing range (x-axis) and estimation acuity (y-axis) across information capacities (colors). Changes in capacity always change the observed estimation acuity but only sometimes dramatically change the subitizing range.

Because the subitizing range can change dramatically without requiring essentially any change in estimation acuity[1], it may not be altogether surprising that some studies have found that the subitizing range is affected by an attentional manipulation when estimation acuity is not. The relationship between the subitizing range and estimation acuity should only become apparent with substantial changes in capacity — and even then, estimation acuity need not change by a substantial margin. For instance, to increase the subitizing range from 2 to 4 would only require a decrease of the coefficient of variation in estimation from 0.27 to 0.21 (highlighted in Figure S7 by the dashed lines). This level of change seems insubstantial relative

[1]One curious thing to note is that when the subitizing capacity changes, the observed estimation acuity actually very slightly *decreases*. This is because numerosities very near the subitizing range tend to have slightly higher acuity than larger numerosities, but when the subitizing range increases to encompass that numerosity, it is no longer counted towards the average estimation acuity.

to the change in subitizing range — and may even be hard to detect without high statistical power — but does not imply that the two phenomena are unrelated.

Small vs. large quantity estimation in children

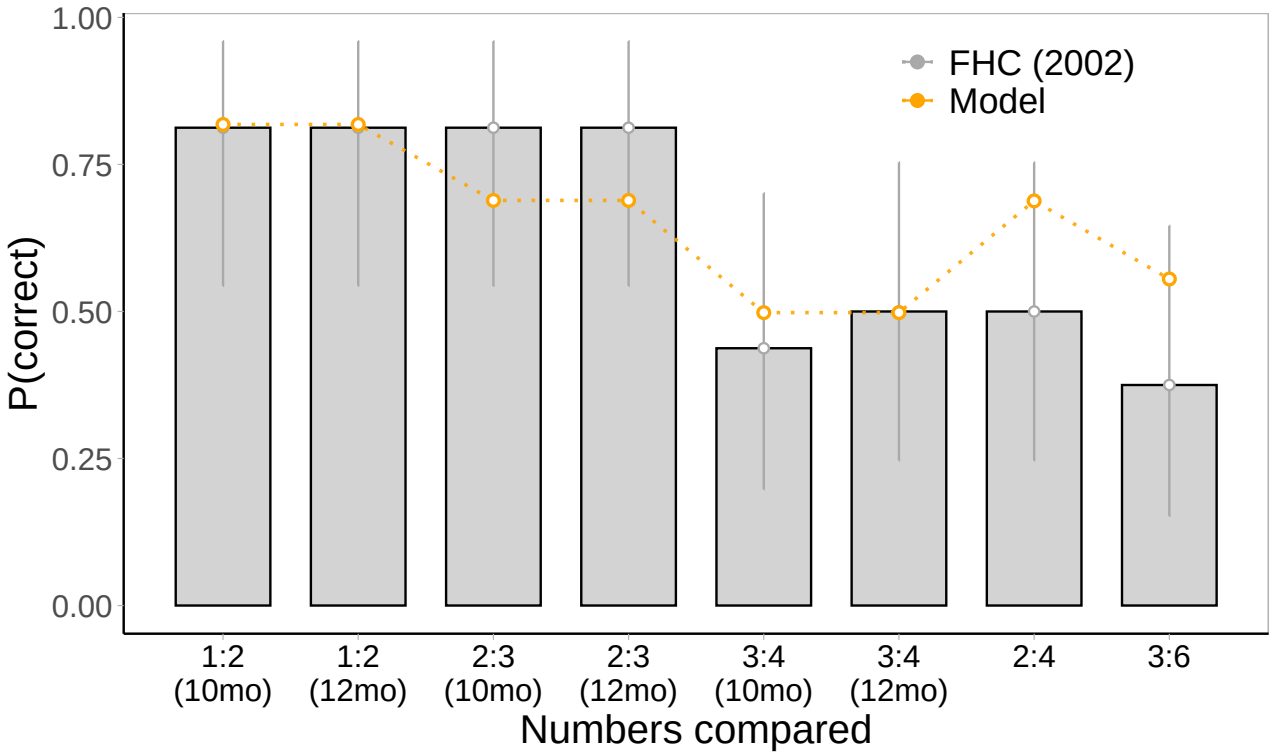


Fig. S8: Data on infants' success in comparing quantities from Feigenson et al. (2002) (gray bars with 95% CI) plotted against our model fit to their data (orange points).

Several studies using habituation and manual search paradigms have found that children succeed in comparing two small quantities (1-3) or two larger quantities (4+), but fail to compare a small set to a larger set (Feigenson & Carey, 2003; Feigenson et al., 2002; Lipton & Spelke, 2004; Xu, 2003). This has been taken as evidence of two separate systems: the parallel individuation system for exactly tracking small sets and an approximate system for inexactly representing magnitudes. This finding is not obviously compatible with our account, though our model does have an individuation component that precedes estimation, so one potential explanation is that children do not automatically map small sets of objects to quantities as adults do — and once they observe a new set of objects, they can no longer perform such a mapping if it is required and hence cannot discriminate small from large sets.

However, we note that our model — as it is, without any adjustment to explain these findings

— is not actually incompatible with at least some of the published data that have been used to support the claim of two independent systems. To illustrate this, we fit our model to the data reported in Feigenson et al. (2002), in which 10-12mo infants were shown two boxes in which they observed the researcher placing different numbers of crackers. The researchers then recorded which box the infants first attempted to search, assuming they would reach toward the one with the greater number of crackers if they could. They found that infants reached toward the box with more crackers given 1 versus 2 crackers and 2 versus 3 crackers, but not on 3 versus 4 crackers, 2 versus 4 crackers or 3 versus 6 crackers. We recovered parameters of 12.9 bits of information, loss function α of 0.21, and guessing rate p_g of 0.36. Figure 8 shows the data from Feigenson et al. (2002) re-plotted against the predictions of our model. In each case, the model predictions fall within the 95% confidence interval (all differences were insignificant in binomial tests, ps > 0.1).

References

Feigenson, L., & Carey, S. (2003). Tracking individuals via object-files: Evidence from infants' manual search. *Developmental Science*, *6*(5), 568–584.

Feigenson, L., Carey, S., & Hauser, M. (2002). The representations underlying infants' choice of more: Object files versus analog magnitudes. *Psychological science*, *13*(2), 150–156.

Lipton, J. S., & Spelke, E. S. (2004). Discrimination of large and small numerosities by human infants. *Infancy*, *5*(3), 271–290.

Xu, F. (2003). Numerosity discrimination in infants: Evidence for two systems of representations. *Cognition*, *89*(1), B15–B25.