nature human behaviour

Article

https://doi.org/10.1038/s41562-023-01602-z

A quantitative model of ensemble perception as summed activation in feature space

In the format provided by the authors and unedited

Supplementary Information

Supplementary Discussion.

Assessing model generalizability. Our core comparison of the TCC ensemble models involves estimating a signal-to-noise ratio from standard working memory tasks for individual items, substituting these parameters into the ensemble models, and predicting data on the ensemble task. This generalization approach involves evaluating each model's capacity to predict unseen data on a new task, which contrasts to many mainstream modeling approaches that simply evaluate models based on their ability to fit data (Yarkoni & Westfall, 2017). While we report results from model fitting, and find that the best performing TCC ensemble models do a good job of capturing key qualitative and quantitative trends in the data (see Supplement), evaluating models based on their capacity to predict data in new tasks meets one of the highest technical and conceptual benchmarks for model comparison. We summarize our approach here to explain how it moves beyond many standard model comparison approaches in the psychological sciences.

On the technical side, assessing models based on their capacity to predict data on a novel task, rather than simply fit data, circumvents a key issue in model evaluation, which is ensuring that a model does not *overfit* the data. Overfitting means that a given model fits nuisance variance, such as sampling error, in addition to, or instead of substantive variance that arises from changes in the latent cognitive processes of interest (e.g., Myung, Tang, & Pitt, 2009). Overly flexible models will tend to overfit data and, in principle, a model can fit a sample of data perfectly even if it is not the true generative model. This is one reason why several researchers have pointed out that assessing models solely based on their fit to data may provide little to no insight into their capacity to capture the cognitive constructs psychologists seek to measure and study (e.g., Roberts & Pashler, 2000). It is also well-known that standard metrics of model fit, which are designed to penalize overly-flexible models based on their number of parameters (e.g., Akaike and Bayesian Information Criteria) may be inadequate because a model's flexibility is jointly determined by its number of parameters and functional form. For this reason, quantitative researchers have promoted evaluating a model's ability to predict unseen data, rather than fit data (e.g., Busemeyer & Wang, 2000; Lee, 2011; Yarkoni & Westfall, 2017), because models that overfit data will tend to do poorly at generalizing to new samples of data, in which error variance will also differ. Thus, evaluating a model's prediction provides a sound

technical way of comparing models while taking into account any potential differences in their flexibility.

On the conceptual side, evaluating a model's ability to make zero-free-parameter predictions on new tasks provides insight into whether the model is capturing key cognitive invariants – or stable latent processes. This conceptual issue has received much recent focus (e.g., Yarkoni, 2022) because psychology researchers need benchmarks for determining whether their theories and models have the capacity to generalize to more real-world tasks. One such major benchmark is to evaluate whether a model has the ability to yield parameters that generalize in a principled way across different laboratory tasks. Evidence that a model has the capacity to generalize in this way provides strong convergent support for the view that it also has potential to capture ecologically meaningful cognitive processes, as opposed to idiosyncratic processes elicited by a specific laboratory task.

Finally, in the current context, our linking approach allows us to directly test central hypotheses regarding how processing of individual items connects to processing of ensembles. That is, we estimate a parameter in a working-memory task and substitute it into the ensemble models to predict the ensemble data (and vice versa). In this way, we test whether patterns of activations elicited by individual items are also involved in processing of ensembles, as well as how they are pooled in the process of ensemble computation. A very similar linking approach has been used in the neural cognitive literature to link neural and cognitive models of processes (Turner et al., 2017; Wong & Wang, 2006), in which researchers substitute measures of neural data – such as patterns of activation in a given set of neurons – in place of free parameters in cognitive models to formally test specific linking propositions. In the current article, we use this approach to directly link processes across different cognitive tasks. Moreover, we use it to directly test fundamental predictions in the ensemble literature about how ensemble representations are constructed. For instance, our comparison of the Perceptual Summation and Post-perceptual Summation models is only meaningful in the context of this generalization approach because they differ only in their prediction of the time-course of ensemble extraction. That is, unlike the Post-perceptual Summation model, the Perceptual Summation model postulates that activations of individual items are pooled before as opposed to after post-perceptual noise accrues, yielding a higher signal-to-noise ratio of the final ensemble representation. We test this prediction directly by obtaining an independent estimate of the signal-to-noise ratio in a working-memory task and substituting it into the models based on these hypotheses. Through this lens, our

generalization approach can also be seen as a method for testing specific linking propositions, but doing so across different cognitive processes rather than neural and cognitive processes.

Alternative variants of the Automatic Averaging model. In the main text we present one variant of the Automatic Averaging TCC ensemble model that postulates that the signal-to-noise ratio of the ensemble representation is the same as it is for a single item in a six- or eight-item memory load condition. However, because the Automatic Averaging ensemble theory has never been formulated as a general computational model, many aspects of this theory, including the processes that underpin automatic averaging, are underspecified. This makes it difficult to make strong assumptions about the signal-to-noise ratio for this model. To address this limitation, we consider a broader range of assumptions about the signal-to-noise ratio for this model. In particular, we also consider the assumptions that the signal-to-noise ratio of the ensemble representation is the same as it is for storing a single item (in a one-item memory load condition). This assumption follows if people automatically extract the average at an early perceptual stage, rather than a later processing stage. To this end, we rescaled the signal-to-noise ratio to the predicted signal-to-noise ratio if people were maintaining only a single item in memory, under the assumption that the signal-to-noise ratio changes as a function of memory load linearly or via a power law (Schurgin, et al., 2020). We found that the power law assumption yielded higher predictive accuracy and focus on this analysis for ease of exposition. Here we also emphasize that we rescaled the signal-to-noise ratio as opposed to measuring it in another working-memory task (for one item) because we need to keep experimental conditions matched across the ensemble and working-memory tasks. In Experiments 1, 2, 3 and 4, we found that the Perceptual Summation model outperformed the Automatic Averaging model (Experiment 1: Perceptual Summation $\overline{X} = 708$ and Automatic Averaging $\overline{X} = 726;\ t(49) = 5.61,\ p < .0001$; Experiment 2: Perceptual Summation $\overline{X} = 1,056$ and Automatic Averaging $\overline{X} = 1,111;\ t(49) = 6.49,\ p < .0001$; Experiment 3: Perceptual Summation $\overline{X} = 766$ and Automatic Averaging $\overline{X} = 784;\ t(49) = 6.6,\ p < .0001$; Experiment 4: Perceptual Summation $\overline{X} = 1,099$ and Automatic Averaging $\overline{X} = 1,123;\ t(49) = 7.53,\ p < .0001$). Together, these analyses align with the model fitting and reverse analysis results because they indicate that the Automatic Averaging model performs worse than the Perceptual Summation model not simply because of our assumptions about the models' signal-to-noise ratio, but because of their difference in functional form, and provides converging support for the Perceptual Summation model.

TCC ensemble model fits. We report two measures of goodness of fit for the TCC ensemble models. The fits of the set of non-TCC ensemble models we consider are identical to the results we report in the main text. The first measure of goodness of fit is R^2, the proportion of variance captured by the model, where $R^2 = 1$ means that a given model captures all of the variance in the data. Note that, in general, observing this upper bound in empirical data is not desirable. This is because each sample of data is necessarily corrupted by sampling error, and if a model perfectly captures all of the variance in a sample of data, this also likely means that it is overfitting error variance in addition to substantive variance – that is, the model may be overly flexible. We discuss relevant issues of model flexibility and some limitations of model assessment solely based on 'goodness of fit' in the Methods section. In the current context, R^2 is also a heuristic summary of model fit, which we report to provide a cursory sense of how well the models accommodate the data. This is because our modeling involves fitting models to distributions of data, and R^2 will vary depending on how we bin the data to approximate these distributions. For example, the possible number of bins in our data is 360 – however, with limited samples of trial-data it is not possible to perfectly approximate this distribution in principle, and R^2 would be very low. At the other extreme, if we used two bins to represent our data and assess model fit, then we would obfuscate all meaningful variation and get an inflated measure of R^2. Note that this limitation is not unique to our design or our modeling – it is simply a consequence of analyzing empirical distributions, which is standard in the visual working-memory literature. To address this problem for a cursory analysis, we binned data using the Freedman Diaconis criterion (Freedman & Diaconis, 1981), which is an approximation designed to minimize the integral between the empirical histogram and the theoretical probability distribution. Using this criterion, we found that 15 bins provided the best approximation to the theoretical distribution, in line with prior work (Zhang & Luck, 2008).

Because R^2 is a cursory measure of model fit, we report the aggregate values across participants, and base primary comparisons on the log likelihood (LL), which is an unbounded measure that is not subject to these limitations.

As described in the main text, the Perceptual and Post-perceptual Summation models differ in their predictions of the signal-to-noise ratio of the ensemble representation. Therefore, if the

signal-to-noise ratio parameter d' is free to vary without regard to how well it can explain the separate working memory task, they will yield identical fits to data (albeit with different parameter values; see the Reverse Inference analysis, in which we find that the Perceptual Summation model outperforms the Post-perceptual Summation model when predicting working-memory data). For simplicity, we refer to fits between the Perceptual Summation and Automatic Averaging model. Figure 1 shows model fits for each of the five experiments.

In Experiment 1 we used color and manipulated memory load. There we found an average $R^2 = .934$ and $R^2 = .903$ for the Perceptual Summation and Automatic Averaging models, respectively. The Perceptual Summation ($\overline{X} = 698;\ SEM = 4.7$) model also provided better fit to data than the Automatic Averaging ($\overline{X} = 705;\ SEM = 4.8$) model based on LL ($t(49) = 6.63,\ p < 1e-07$).

In Experiment 2 we used color and manipulated the range of colors. There we found an average $R^2 = .936$ and $R^2 = .873$ for the Perceptual Summation and Automatic Averaging models, respectively. The Perceptual Summation ($\overline{X} = 1,041;\ SEM = 7.6$) model also provided better fit to data than the Automatic Averaging ($\overline{X} = 1,054;\ SEM = 8.4$) model based on LL ($t(49) = 6.69,\ p < 1e-07$).

In Experiment 3 we used shape and manipulated memory load. There we found an average $R^2 = .884$ and $R^2 = .824$ for the Perceptual Summation and Automatic Averaging models, respectively. The Perceptual Summation ($\overline{X} = 746;\ SEM = 7.0$) model also provided better fit to data than the Automatic Averaging ($\overline{X} = 754;\ SEM = 7.2$) model based on LL ($t(49) = 7.47,\ p < 1e-08$).

In Experiment 4 we used shapes and manipulated the range of shapes. There we found an average $R^2 = .897$ and $R^2 = .866$ for the Perceptual Summation and Automatic Averaging models, respectively. The Perceptual Summation ($\overline{X} = 1,084;\ SEM = 11.1$) model also provided better fit to data than the Automatic Averaging ($\overline{X} = 1,090;\ SEM = 11.3$) model based on LL ($t(49) = 4.49,\ p < 1e-04$).

In Experiment 5 we tested participants on the color of sequentially presented objects. There we found an average $R^2 = .912$ and $R^2 = .679$ for the Recency Perceptual Summation and

Automatic Averaging models, respectively. The Recency Perceptual Summation ($\overline{X} = 453;\ SEM = 5.3$) model also provided better fit to data than the Automatic Averaging ($\overline{X} = 473;\ SEM = 5.3$) model based on LL ($t(49) = 9.57,\ p < 1e - 12$).

Together, we find that each of the TCC ensemble models does an adequate job of distribution of continuous report data in ensemble memory tasks, with the Perceptual and Post-perceptual Summation models slightly but systematically outperforming the Automatic Averaging model.

Reverse Inference. For our central analyses, we compare each ensemble model's capacity to predict the ensemble data. In general, we find that ensemble models from the TCC framework, which link memory for individual items to memory for ensembles, outperform a set of alternative point estimate models. These TCC models embody core predictions in the ensemble literature regarding how ensemble processes unfold over time, how ensembles relate to processing of individual items and the nature of ensemble representations. For this subset of best performing TCC models, we can also ask the following: How do the models compare in prediction when we reverse inference, that is, when we attempt to predict performance in the working memory tasks when estimating the signal-to-noise of each memory representation from the ensemble data with the TCC ensemble models?

Answering this question yields two insights. First, it elucidates whether our modeling results that the Perceptual Summation model is the best performing model are robust across different ways of predicting the data. Accordingly, this also provides insight into whether the processing assumptions of this model are tenable. For instance, if we find that the Perceptual Summation model compares poorly in across-task prediction to the Post-Perceptual Summation and/or the Automatic Averaging model this would be problematic because it may indicate that it does not do a superior job of capturing latent cognitive processes in the ensemble task. This might suggest that this model's capacity to predict ensemble data in our core analyses is due to a limitation of the model, such as that its functional form does an adequate job of approximating distributions of data in ensemble tasks, but is equally compatible with a range of assumptions regarding how the signal-to-noise ratio for individual items links to memory for ensembles. Although our comparison of the Perceptual Summation and Post-perceptual Summation model speaks against this interpretation, the reverse inference analysis provides an alternative and direct test of it. Conversely, if we find that the Perceptual Summation model is the best performing model in the reverse inference analysis, this would indicate that the Perceptual

Summation model captures cognitive invariants in ensemble tasks, and help validate our framework for linking memory for individual items to ensembles more broadly.

To summarize, we found the same patterns of results with the reverse inference analysis as we did with the core analysis. These results are summarized for all five experiments in Table 1. We show predictions of working memory data for each of the TCC ensemble models in Figure 2. As before, we found that the Perceptual Summation model outperforms each of the alternative TCC models in prediction. That is, we found that we could fit the Perceptual Summation model to the ensemble data, use it to estimate the signal-to-noise ratio for each individual item representation, and then substitute this signal-to-noise ratio into the TCC model for individual items (Equations 2 and 1 in the main text, respectively) to predict performance the working memory data. This analysis helps validate our conclusion that the assumptions of the Perceptual Summation are tenable, and corroborates our framework for linking performance between memory tasks for ensembles and individual items.

Predictions of other phenomena in ensemble literature. The goal of our modeling work was to develop and test a principled, generalizable computational model of ensemble processing that could account for a wide range of phenomena across different task demands. For this reason, we focused our empirical work on a range of mainstream ensemble tasks and manipulations, such as memory load, range of feature values, and presentation format, that yield relatively robust effects on ensemble memory. In this section we report simulations from the Perceptual Summation model to demonstrate how the model could account for a wider range of phenomena, with a specific focus on ensemble tasks that either yield inconsistent results, such as ensemble tasks with outliers, or are not typically used in the ensemble literature, but may elucidate theoretically interesting boundary conditions of the model, such as ensemble tasks with randomly sampled stimuli.

The first set of simulations focuses on the potential effects of outliers on ensemble processing, that is, a single item that has a feature value that is very far from the other items in the ensemble array. As shown in the left panel of Figure 3A, the Perceptual Summation model straightforwardly predicts discounting of outliers, which has often, though not always, been reported in the literature (e.g., Haberman & Whitney, 2010; Li et al., 2017). In these simulations we assumed that the outlier stimulus is weighted in the same way, that is, it has the same signal-to-noise ratio, as other stimuli in the ensemble array. However, even though we make assume equally-weighted representations, the Perceptual Summation model predicts that the

outlier is ‘discounted’ simply because, after pooling via summation, the distributed patterns of activation elicited by other items in the ensemble array collectively outweigh the levels of activation elicited by the outlier stimulus, and, therefore, the activation of feature values closest to the with maximum, is almost entirely unaffected by the outlier. Thus, the model naturally predicts that outliers should be discounted in ensemble tasks even when they are processed in the same way as other items in the memory array.

However, in principle, the model can also be accommodated to capture increased weighting of the outlier as shown in the right panel of Figure 3A. Through the lens of the TCC framework, such increased weighting of the outlier would entail a higher signal-to-noise ratio (larger d' estimate) for the outlier relative to other stimuli in the ensemble array. There are likely situations where this occurs, such as at sufficiently long encoding times, or when the task demands are set up in such a way as to encourage increased attentional processing to salient items that do not fit the display (e.g., as modeled by Brady & Tenenbaum, 2013). The general propensity of models like the one proposed here to discount outliers may thus sometimes be counteracted by people’s tendency to attend to and encode outliers more strongly (such as in conditions where a stimulus ‘pops-out’, e.g., Treisman & Gelade, 1980), perhaps explaining why there are mixed results in the literature about whether outliers are discounted (e.g., Cant & Xu, 2020; Khayat & Hochstein, 2018; Hochstein, Pavlovskaya, Bonneh, & Soroker, 2018).

To quantitatively test this set of predictions about outlier discounting, researchers would need to obtain a separate d' estimate for the relatively homogenous items and the outlier stimulus in a visual working memory task. This is not straightforward to do, however, because participants may use high-level strategies in such a task, relying on the other items for items in the homogenous set (e.g., Nassar et al. 2018). In addition, we note that the model is not designed to capture higher-level, decision stage processes that people may use to incorporate the outlier into the ensemble if it is over-weighted relative to other items. For instance, it is possible that if people become aware there is an outlier, they do not simply read out the patterns of activation elicited by the ensemble array, but use additional compensatory strategies to calibrate their responses so they approximate the average more accurately (e.g., explicitly shift their responses toward the average of the homogenous set of items). Proposing a comprehensive theory of outlier processing is outside the scope of our work, but our simulations point to reasons why outlier tasks may vary in how they affect ensemble extraction.

Another phenomenon which we did not focus on in the current empirical work is that ensemble representations may sometimes become more precise with an increase in set size (e.g., Allik et al., 2013; Baek & Chong, 2020; Haberman & Whitney, 2010; Lee et al., 2016; Parkes, Lund, Angelucci, Solomon, & Morgan, 2001), and sometimes not (e.g., Ji & Pourtois, 2018). As shown in Figure 3B, the Perceptual Summation model predicts such increased precision effects under some experimental conditions but not others. Specifically, under conditions where the range of feature values is the same across different set sizes, the Perceptual Summation model predicts a relative increase in precision for higher set size (compare leftmost panel with rightmost panel in Figure 3B). However, under conditions where the step size between feature values is fixed and the range varies — as in our experiments — the Perceptual Summation model predicts a small drop in precision for higher set sizes even for the same values of d' (compare the leftmost panel to the center panel in Figure 3B). We believe this may account for some of the differences in the literature with regard to whether set size improves or does not improve ensemble performance.

Thus, by the nature of the summation process and representation of items and ensembles as distributions over feature space, the model predicts differential effects of memory load on the precision of ensemble representations, depending on the distribution of feature values of to-be-remembered stimuli. Notably, our empirical results and simulations provide a parsimonious account for a pattern of results that was previously used to motivate a specific subsampling account of ensemble processing. Specifically, Marchant, Simons and Fockert (2013) found that ensemble performance decreased when the number of items and their heterogeneity increased. To account for these inconsistent effects of memory load on ensemble processing, the authors proposed that people selectively subsample items in the ensemble array, store them and in visual working memory, and use these working memory representations of individual items to compute an average at probe (similar to the prediction of the of Post-perceptual Summation model). Our simulations indicate that these differential effects of set size on ensemble processing can be predicted in a straightforward fashion by the signal detection based Perceptual Summation model without any appeal to subsampling.

Finally, we consider how differences in distributions of feature values affect ensemble reports for a fixed memory set size. In particular, we examine what the model predicts when feature values are drawn randomly, or in clusters from a single feature continuum. We did not perform such studies in our empirical work because a basic assumption in ensemble processing in experimental and ecological conditions is that ensemble representations are computed when

there are redundancies across items in the external environment. Random sampling of feature values does not ensure such overlap, and makes it conceptually unclear how people would extract an ensemble representation under such conditions.

Nevertheless, we demonstrate that the model would make conceptually plausible predictions for the distribution of activation patterns elicited by the ensemble array for distributions of feature values that are not drawn from just a single (e.g., uniform) cluster. This could potentially be used as the basis for 'chunking' (or the formation of 'clusters') in visual working memory displays (e.g., Son et al. 2020), providing a unifying principle for such chunking strategies and for ensemble perception.

In particular, as shown in Figure 3C, if there are multiple (six) objects in the ensemble array, the predicted patterns of activation of the ensemble representation will depend on how feature values are distributed. If the distribution is close to uniform (leftmost panel of Figure 3C), the predicted patterns of activation will also approach uniformity, *not* giving rise to well-represented 'chunks.' If there is some clustering in the ensemble array (center panel of Figure 3C), the predicted patterns of activation of the ensemble will show slight clustering, particularly for the items that are similar to each other. Finally, the rightmost panel of Figure 3C shows that if stimuli values are constrained to be sampled in clusters, such 'ensembles' would elicit bumps of activation analogous to 'chunks.' While we do not investigate whether participants use such a strategy to form chunks in more heterogeneous displays, the connection between the ensemble model and a potential model of chunking is a useful avenue for future work.

Connections to other quantitative models of ensembles. In this section we review how the Perceptual Summation model, as well as our general modeling approach compares and contrasts with existing ensemble models. We took a top-down approach towards model development, meaning that we used an existing framework of memory for individual items to develop, a priori, a memory for ensembles. We contrast this top-down approach to a task-driven model development in which models are developed to fit data to a specific task. To our knowledge, the majority of quantitative models of ensemble processes fall within the scope of such task-specific models. For instance, Solomon's (2010) ensemble model was developed to fit data for orientation ensemble tasks, the distributed attention model (Baek and Chong, 2020) was developed to fit data from an ensemble task for size judgments, and Tong, Dube, and Sekuler (2019) ensemble model was developed to fit data for sequential ensemble tasks. Although there is certainly potential for extending these models, it remains unclear how well

they would generalize to a wider range of ensemble tasks and stimuli spaces. Moreover, in their current form, none of these models make high-precision predictions of performance by capturing the full distribution of memory errors in a continuous report task, nor generalize from individual item memory to ensemble memory.

Absent a quantitative comparison, we note informally that there is conceptual overlap between these models and the Perceptual Summation model. For instance, like the Baek and Chong (2020) and Solomon (2010) models, the Perceptual Summation model is grounded in signal detection theory, and postulates that ensemble representations can be corrupted at different stages of processing (here, a perceptual pre-summation stage and memory post-summation stage). In addition, we leveraged the Tong et al., (2019) model to develop the Recency Perceptual Summation model and, as such, our model is also aligned with the view that ensemble memory representations in sequential presentation tasks are subject to exponential decay. Furthermore, like the Tong et al., (2019) our ensemble model is broadly consistent with prototype models because it posits an abstraction (via summation) over representations of individual items. However, our model also postulates that the decision variable in the ensemble task reflects a sum of similarity signals from all individual item representations and, therefore, also inherits some assumptions from exemplar models (e.g., Kahana & Sekuler, 2002; Nosofsky, 1988). Through this lens, our model fits with more nuanced views (e.g., Lee & Vanpaemel, 2008; Ross & Makin, 1999; Tong et al., 2019) in which prototype and exemplar representations can be seen as endpoints on a continuum, as opposed to mutually exclusive classes of representations.

Signal Detection Theory based modeling approach. In this section, we address a few conceptual questions related to Signal Detection Theory and parameter generalization. First, we note that every computational cognitive model makes simplifying assumptions and the appropriate level of specificity depends on the modeling goals (e.g., Farrel & Lewandowsky, 2018). Our modeling goal was to develop a model of ensembles perception that speaks to core issues in the ensemble literature, such as whether ensemble representations are probabilistic, and how noise accrues over them, as well as to formally connect a model of visual working memory for individual items and ensembles. As a starting point towards this linking analysis, we used a quantitative measure of the signal-to-noise ratio in the visual working memory task for individual items. That is, we quantified memory ‘strength’ for each item via its signal-to-noise ratio, rather than specifying its entire representational format, and examined whether it can be substituted into a process-level ensemble model to predict individual differences in performance

on the ensemble task. Naturally, future work can focus on creating a more detailed process-level account by unpacking the processes measured by the signal-to-noise ratio. However, in light of our goal of predicting performance across different memory tasks, more parsimonious models provide a useful springboard for generalization.

One concern may be that d' is not only quantifying the signal-to-noise ratio due to memory related processes, but also other non-memory related processes, such as individual differences in task-engagement. Accordingly, the finding that d' predicts individual differences in performance across tasks, may reflect, in part, its ability to capture individual differences in third variables such as how conscientiously people approach the task. We certainly do not deny that part of the variance captured by d' may reflect individual differences in other global factors such as task engagement, which undoubtedly have downstream effects and explain inter-individual differences in memory performance in *all* memory tasks. However, our results from intra-individual analyses (within-subject conditions) — for example, how both item memory and ensemble memory change as a function of range or set size — are not possible to explain by appealing solely to overall differences in motivation across participants. In addition, data from the sequential paradigm (Experiment 5) provide strong evidence against the view that the predictive value of d' reflects solely on such across-subject factors, and not individual differences in memory-related processes. This is because in Experiment 5 we show strong support for the view that our models and generalization approach also captures intra-individual variance in performance (as captured by changes in d' as a function of serial position), that is, memory changes as a function of the serial position of the stimulus within the same person. Such within-task predictions would be impossible to make if d' only captured non-memory related processes.

Another point relevant to our Signal Detection-based modeling is that aspects of the signal detection based TCC model of memory were recently questioned by Oberuaer (2022) and Tomic and Bays (2022). We note here, however, that both the analysis of Oberauer (2022) and Tomic and Bays (2022) only assessed models based on their ‘goodness of fit’ not their capacity to generalize across task structures. As we discuss here, and has been discussed extensively over the last couple of decades (e.g., Roberts & Pashler, 2000), the ‘goodness of fit’ of a given model can provide little to no insight into its capacity to measure or explain cognitive constructs. Indeed, in more recent work we used a critical test and found evidence against a conclusion by Oberauer (2022) (Robinson, DeStefano, Brady, & Vul, 2021). Moreover, unlike the TCC model, neither of the best performing models identified by Oberauer (2022) or Tomic and Bays (2022)

have been shown to make principled, zero-free-parameter generalizations across task structures. Based on this criterion, this evidence does not provide strong evidence against the core assumptions of TCC.

Finally, we note that we do not instantiate models in which participants ‘sub-sample’ items from the array (Myczek & Simons, 2008). This subsampling view entails that participants fail to process the remaining subset of items at an early encoding stage, a view compatible with item-based capacity visual working memory models (e.g.,Cowan, 2001; Zhang & Luck, 2008) that postulate all-or-none memory states (e.g., Kellen & Klauer, 2015). However, such a view is incompatible with much current evidence for, signal detection theory-based continuous resource models of memory (e.g., Schurgin, et al., 2020; Robinson et al. 2020; van den Berg et al., 2014; Williams, Robinson, Schurgin, Wixted, Brady, 2022). Relatedly, this view is also incompatible with most modern accounts of visual working-memory (e.g., van den Berg et al. 2012; Bays, 2014; Schneegans et al. 2020; Schurgin et al. 2020) according to which all memory representations are noisy, rather than all-or-none. If future work reveals that, unlike visual working-memory processes, ensemble perception does involve complete information loss about some subset of items, such models could be tested in the current framework; for instance, with a mixture of d' values for different items, with some set to zero and others set to be greater than zero.

Supplemental References

Allik, J., Toom, M., Raidvee, A., Averin, K., & Kreegipuu, K. (2013). An almost general theory of mean size perception. *Vision Research*, 83, 25–39. https://doi.org/10.1016/j.visres.2013.02.018

Baek, J., & Chong, S. C. (2020). Ensemble perception and focused attention: Two different modes of visual processing to cope with limited capacity. *Psychonomic Bulletin & Review*, 27(4), 602–606. https://doi.org/10.3758/s13423-020-01718-7

Bays, P. M. (2014). Noise in neural populations accounts for errors in working memory. *Journal of Neuroscience,* 34(10), 3632–3645. doi:10.1523/JNEUROSCI.3204-13.2014

Brady, T. F., & Tenenbaum, J. B. (2013). A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological Review*, 120(1), 85–109. https://doi.org/10.1037/a0030779

Cant, J. S., & Xu, Y. (2020). One bad apple spoils the whole bushel: The neural basis of outlier processing. *NeuroImage*, 211, 116629. doi.org/10.1016/j.neuroimage.2020.116629

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, *24*(1), 87-114.

Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. Cambridge University Press.Li et al., 2017

Freedman, D., & Diaconis, P. (1981). On the histogram as a density estimator: L2 theory. *Probability Theory and Related Fields*. 57 (4): 453–476.

Haberman, J., & Whitney, D. (2011). Efficient summary statistical representation when change localization fails. *Psychonomic Bulletin & Review*, 18(5), 855–859.

https://doi.org/10.3758/s13423-011-0125-6

Hochstein, S., Pavlovskaya, M., Bonneh, Y. S., & Soroker, N. (2018). Comparing set summary statistics and outlier pop out in vision. *Journal of Vision*, 18(13), 12. https://doi.org/10.1167/18.13.12

Ji, L., & Pourtois, G. (2018). Capacity limitations to extract the mean emotion from multiple facial expressions depend on emotion variance. *Journal of Vision*, 18(10), 610. https://doi.org/10.1167/18.10.610

Kahana, M. J., & Sekuler, R. (2002). Recognizing spatial patterns: a noisy exemplar approach. *Vision Research*, 42(18), 2177–2192. https://doi.org/10.1016/s0042-6989(02)00118-9

Khayat, N., & Hochstein, S. (2018). Perceiving set mean and range: Automaticity and precision. *Journal of Vision*, 18(9), 23. https://doi.org/10.1167/18.9.23

Lee, H., Baek, J., & Chong, S. C. (2016). Perceived magnitude of visual displays: Area, numerosity, and mean size. *Journal of Vision*, 16(3), 1-11. https://doi.org/10.1167/16.3.12

Lee, M. D., & Vanpaemel, W. (2008). Exemplars, Prototypes, Similarities, and Rules in Category Representation: An Example of Hierarchical Bayesian Analysis. *Cognitive Science*, 32(8), 1403–1424. https://doi.org/10.1080/03640210802073697

Li, V., Herce Castañón, S., Solomon, J. A., Vandormael, H., & Summerfield, C. (2017). Robust averaging protects decisions from noise in neural computations. *PLoS Computational Biology*, 13.

Marchant, A. P., Simons, D. J., & de Fockert, J. W. (2013). Ensemble representations: Effects of set size and item heterogeneity on average size perception. *Acta Psychologica*, 142(2), 245–250. https://doi.org/10.1016/j.actpsy.2012.11.002

Myczek, K., & Simons, D. J. (2008). Better than average: Alternatives to statistical summary representations for rapid judgments of average size. *Perception & psychophysics*, *70*(5), 772-788.

Myung, J. I., Tang, Y., & Pitt, M. A. (2009). Evaluation and comparison of computational models. *Methods in enzymology*, *454*, 287-304.

Nassar, M. R., Helmers, J. C., & Frank, M. J. (2018). Chunking as a rational strategy for lossy data compression in visual working memory. *Psychological review*, *125*(4),486.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(1), 87–108. https://doi.org/10.1037/0278-7393.13.1.87

Oberuaer. K, (2022) Measurement models for visual working memory-A factorial model comparison. *Psychological Review.*

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience*, 4(7), 739–744. https://doi.org/10.1038/89532

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological review*, *107*(2), 358.

Robinson, M. , Brady, T., Vul, E., & DeStefano, I. C. (2021, July). Revisiting the connection between the Luce's choice rule and signal detection theory. *Paper accepted at the 2021 annual Conference of Mathematical Psychology.*

Ross, B. H., & Makin, V. S. (1999). Prototype versus exemplar models in cognition. *The nature of cognition*, 205-241.

Schneegans, S., Taylor, R., & Bays, P. M. (2020). Stochastic sampling provides a unifying account of visual working memory limits. *Proceedings of the National Academy of Sciences*, *117*(34), 20959-20968.

Schurgin, M. W., Wixted, J. T., & Brady, T. F. (2020). Psychophysical scaling reveals a unified theory of visual memory strength. *Nature Human Behaviour*, 4(11), 1156–1172. https://doi.org/10.1038/s41562-020-00938-0

Solomon, J. A. (2010). Visual discrimination of orientation statistics in crowded and uncrowded arrays. *Journal of Vision*, 10(14), 1-16. https://doi.org/10.1167/10.14.19

Son, G., Oh, B. I., Kang, M. S., & Chong, S. C. (2020). Similarity-based clusters are representational units of visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46.

Tomić, I., & Bays, P. (2022). Perceptual similarity judgments do not predict the distribution of errors in working memory. *Journal of Experimental Psychology: Learning, Memory and Cognition.*

Tong, K., Dubé, C., & Sekuler, R. (2019). What makes a prototype a prototype? Averaging visual features in a sequence. *Attention, Perception, & Psychophysics*, *81*(6),1962-1978.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1), 97–136. https://doi.org/10.1016/0010-0285(80)90005-5

Turner, B. M., Forstmann, B. U., Love, B. C., Palmeri, T. J., & van Maanen, L. (2017). Approaches to analysis in model-based cognitive neuroscience. *Journal of Mathematical Psychology*, 76, 65–79. https://doi.org/10.1016/j.jmp.2016.01.001

Van den Berg, R., Shin, H., Chou, W. C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences*, *109*(22), 8780-8785.

Van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological review*, *121*(1), 124.

Williams, J., Brady, T.F., and Störmer, V.S. (in press). Guidance of attention by working memory is a matter of representational fidelity. *Journal of Experimental Psychology: Human Perception and Performance*.

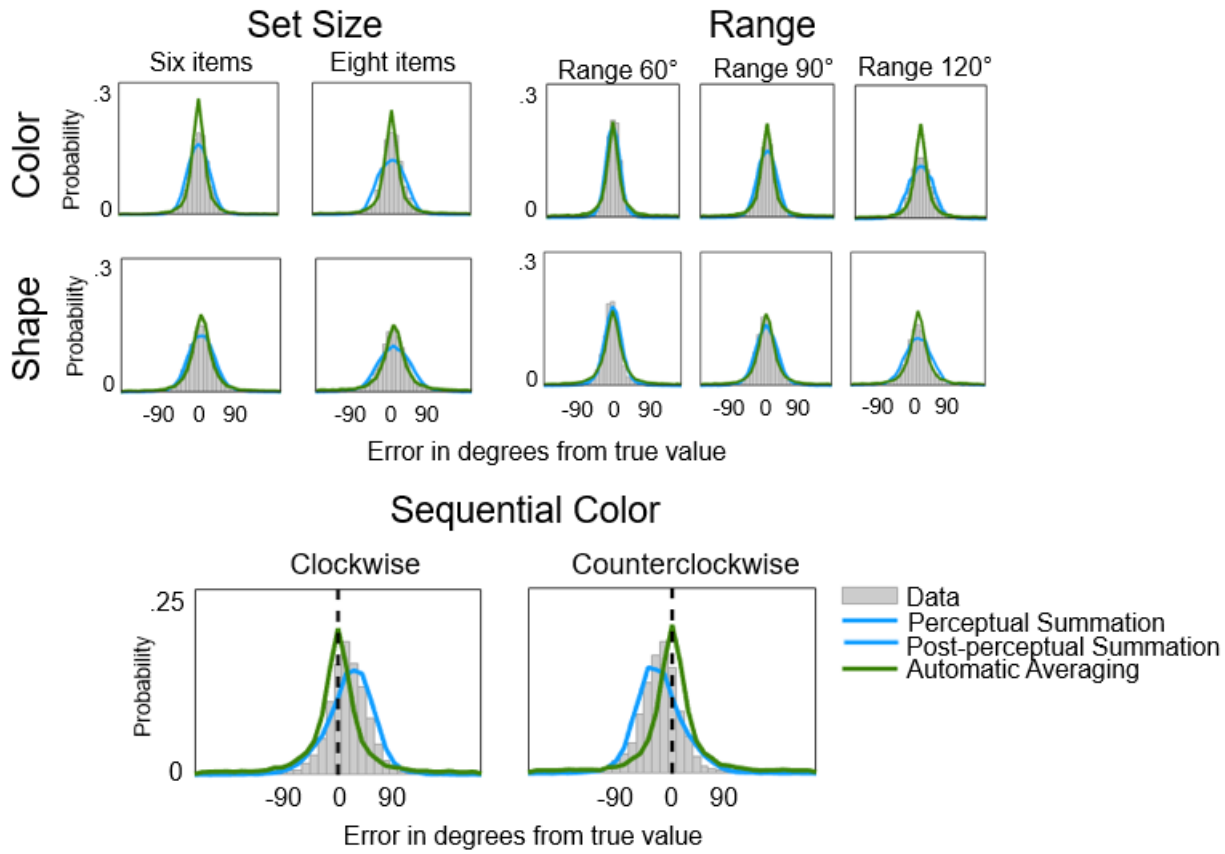
Wong, K.-F., & Wang, X.-J. (2006). A recurrent network mechanism of time integration in perceptual deci-sions. Journal of Neuroscience, 26(4), 1314–1328.

Yarkoni, T. (2021). The generalizability crisis. *Behavioral and Brain Sciences*, 1-37.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100-1122.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, 453(7192), 233–235. https://doi.org/10.1038/nature06860

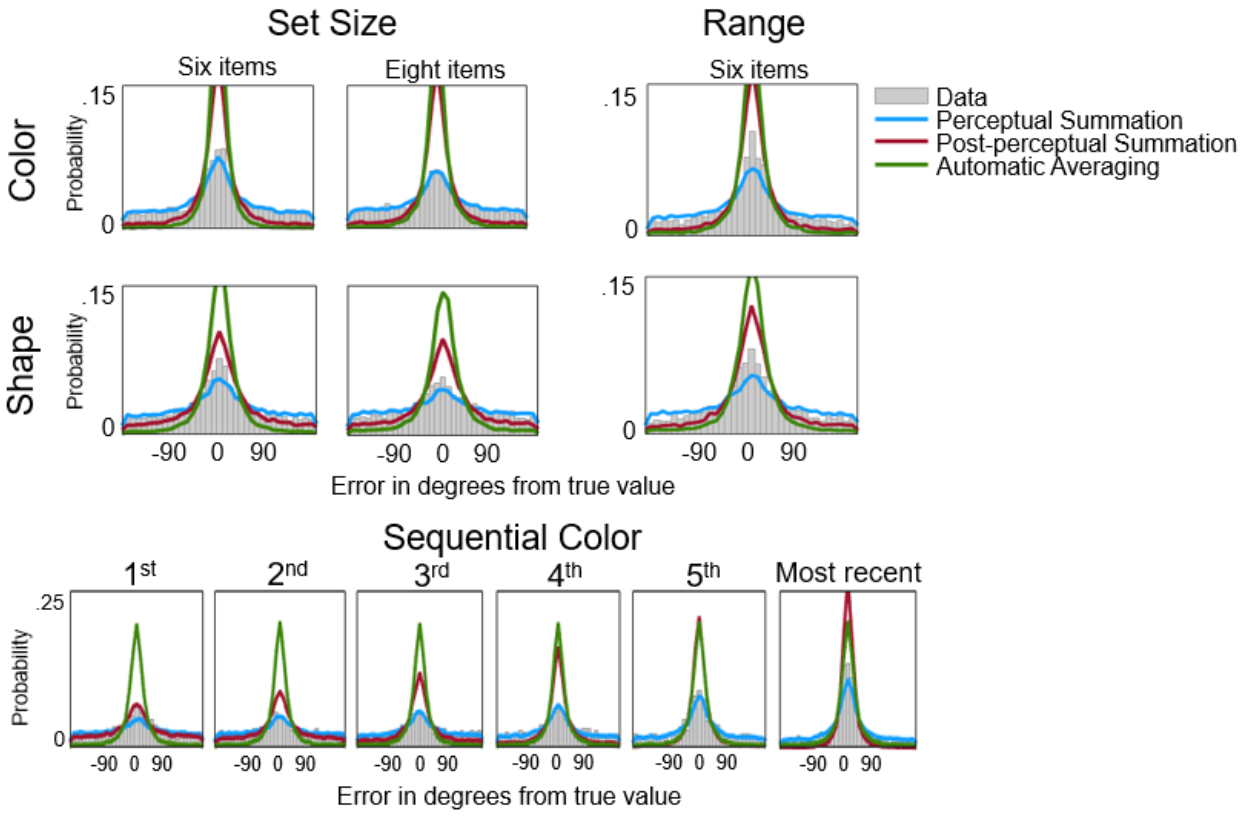
Supplementary Figures



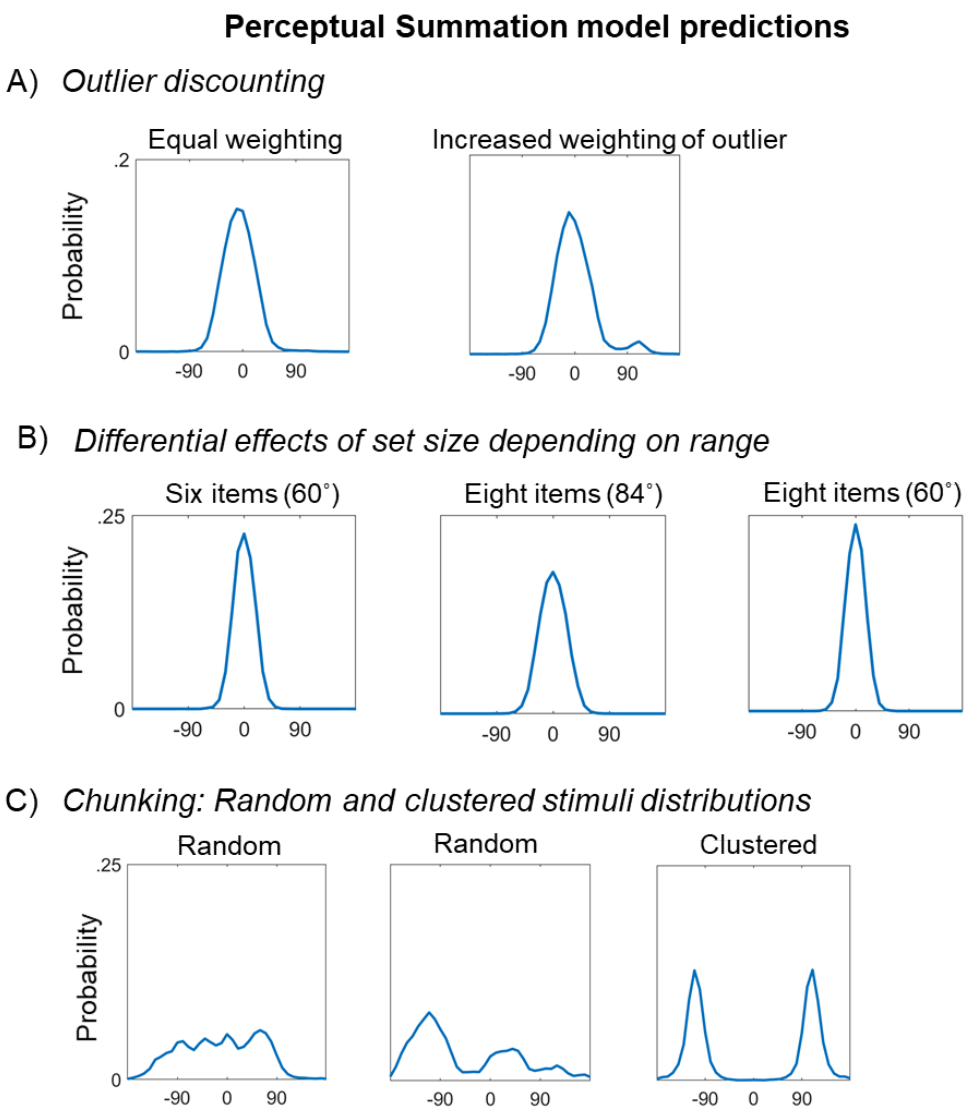
Supplementary Fig. 1 | TCC ensemble model fits to ensemble data*. The top panel shows fits of the TCC ensemble models to the ensemble data in Experiment 1 and 2, (left and right, respectively) in which we tested memory for ensembles of color. The middle panel shows fits of the TCC ensemble models to the ensemble data in Experiment 3 and 4 (left and right, respectively) in which we tested memory for ensembles of shapes. The bottom panel shows fit of the TCC ensemble sequential models to the ensemble data in Experiment 5 in which we tested memory for sequentially presented colored objects.*

	Model	Statistics			
		$t(49)$	$p <$	d_z	CI_L, CI_U
Color: Set Size Perceptual Summation model ($\overline{X} = 867$; $SEM = 2.1$)	Post-perceptual Summation ($\overline{X} = 935$; $SEM = 8.4$)	9.4	$1.0e-11$	1.32	$55, 83$
	Automatic Averaging ($\overline{X} = 965$; $SEM = 6.3$)	20	$1.0e-24$	2.83	$89, 109$
Color: Range Perceptual Summation model ($\overline{X} = 427$; $SEM = 1.5$)	Post-perceptual Summation ($\overline{X} = 443$; $SEM = 4.4$)	4.7	$1.0e-4$	0.67	$9, 23$
	Automatic Averaging ($\overline{X} = 465$; $SEM = 5.2$)	9.6	$1.0e-12$	1.36	$30, 46$
Shape: Set Size Perceptual Summation model ($\overline{X} = 867$; $SEM = 2.1$)	Post-perceptual Summation ($\overline{X} = 889$; $SEM = 4.5$)	5.5	$1.0e-5$	0.78	$14, 31$
	Automatic Averaging ($\overline{X} = 932$; $SEM = 6.3$)	11.7	$1.0e-15$	1.65	$54, 76$
Shape: Range Perceptual Summation model ($\overline{X} = 428$; $SEM = 1.2$)	Post-perceptual Summation ($\overline{X} = 443$; $SEM = 2.7$)	5.3	$1.0e-5$	0.75	$9, 20$
	Automatic Averaging ($\overline{X} = 464$; $SEM = 3.2$)	12.6	$1.0e-16$	1.78	$30, 41$
Color: Sequential Perceptual Summation model ($\overline{X} = 693$; $SEM = 2.5$)	Post-perceptual Summation ($\overline{X} = 736$; $SEM = 5.4$)	9.8	$1.0e-12$	1.39	$34, 52$
	Automatic Averaging ($\overline{X} = 744$; $SEM = 5.3$)	11.7	$1.0e-15$	1.66	$42, 59$

Supplementary Table 1| Summary of descriptive and inferential statistics from all reverse inference TCC ensemble model comparisons across all five experiments.



Supplementary Fig. 2 |Predictions of working memory performance with TCC ensemble models. *Graphical representation of the Perceptual Summation (blue), Post-perceptual summation (red) and Automatic Averaging (green) models' predictions of the working memory data in Experiments 1-5. For this reverse inference analysis, we fit each of the TCC ensemble models to the ensemble data to estimate d', and then substituted these parameter estimates into the TCC working-memory model to predict the working memory data. As before, for this analysis, we found that the Perceptual Summation model outperformed other models in prediction.*



Supplementary Fig. 3 | Predictions of other phenomena from the Perceptual Summation model*. (A) Predicted distribution of errors in ensemble tasks where people have to remember eight items and an outlier is present; in particular, where the distribution of feature values is -53, -38 , -23, -8, 7, 22, and 112 around the true average (in both panels). In the leftmost panel all stimuli are weighted equally and, therefore, have the same* d*', and the outlier — at position 112 — does not significantly skew the resulting response distribution. In the rightmost panel, the outlier is weighted more (*d*' of 2) than the other stimuli (* d*' of 0.82, from Schurgin et al. set size 8). The* d*' value for the outlier was chosen arbitrarily, since there are currently no theories of how much the outlier may be overweighted relative to other items in the array. In this circumstance, the outlier starts to play a role in affecting ensemble performance. (B) Predicted distribution of errors for different set sizes and different stimulus ranges (e.g., different item similarities). The* d*' used for these simulations were taken from real aggregate data from Schurgin et al., 2020 (*d*' for six and eight items is 1.1 and 0.82, respectively). The leftmost panel shows the predicted distribution of errors for six items where all items are equally spaced in feature space, and their range is 60˚. The center panel shows the predicted distribution of errors for eight items where all items are equally spaced and they have the same step-size (12deg) as the six item condition, but a wider range. The rightmost panel shows the predicted distribution of errors for eight items where all items are equally spaced and they have*

a smaller step-size than the six item condition, but the same range. As can be seen, the model predicts different effects of set size depending on how the range of the items is modulated as an effect of set size. (C) Potential generalization of the model to account for chunking. The predicted distribution of errors on the ensemble task for randomly drawn and clustered distribution of stimuli in the ensemble array is shown. The leftmost and center panel show predicted errors for two different simulations in which stimuli feature values are randomly sampled. In the leftmost panel, the distribution of stimuli is -129, -88, -40, 5, 57, and 83 deg from the mean value. In the center panel, the distribution of stimuli is -150, -114, -83, 7 4, 117 deg. Both distributions were obtained by randomly generating integers using the MATLAB built in function randi on the range -180 to 180. The rightmost distribution shows the predicted distribution of errors on the ensemble task if there is clustering in the ensemble array; here the distribution of stimuli values was set to -120, -108, -96, 96, 108 and 120. The model not only provides a model of ensemble perception with a single cluster of items, but seems to provide a promising avenue for investigating how people separate displays into "chunks" as well.