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

Maria M. Robinson* and Timothy F. Brady*

Psychology Department

University of California, San Diego

La Jolla, California, USA

*Email: mrobinosn@ucsd.end; timbrady@ucsd.edu

Abstract

Ensemble perception is a process by which we extract and consolidate redundancies in the environment, getting the ‘gist’ of a set of objects. This ability to summarize complex scenes is thought to underlie our ability to construct robust memory representations, guide attention and classify multiple features or objects. Despite the relevance of ensemble perception to everyday cognition, there are few computational models that provide a formal, process-level account of ensemble perception across a range of experimental conditions and stimuli. In the current work, we take a generalization approach towards developing and testing such a model. According to the proposed theory of ensemble processes, items evoke distributed patterns of familiarity over feature space, and ensemble representations reflect the global sum of these signals across all individual items. We leverage this set of minimal assumptions to formally connect a model of memory for individual items to ensembles. Our approach involves using parameter estimates of performance on a visual memory task for individual items, to make zero-free parameter predictions of inter- and intra- individual differences in performance on an ensemble continuous report task. We compare our ensemble model against a set of alternative models in five experiments. Our top-down modeling approach formally unifies models of memory for individual items and ensembles, and opens a venue for building and comparing models of distinct memory processes and representations. We discuss our model and results in the context of current theories and models of ensemble processing, gist memory and neural population coding.

Human perception and cognition are grounded in a capacity limited system[1,2,3,4]. A basic question across research areas in the behavioral sciences is how people effectively represent an environment that should far exceed their processing capabilities[5]. One widely accepted answer to this question is that perceptual and cognitive systems take advantage of redundancies in the environment by forming a condensed summary or gist[6,7,8]. In the visual domain, the ability of people to extract summaries in this way is commonly referred to as ensemble perception[9].

In a standard laboratory ensemble task, participants are shown a set of stimuli that share properties in a specific feature dimension, such as color, and are instructed to report on their average along that dimension. Figure 1 depicts two example ensemble tasks with colors and shapes. People are remarkably accurate at these tasks, and often notably better at reporting the average of the set than at reporting on any individual item[10]. Extensive empirical and theoretical work suggests that ensemble processing partially underlies our ability to create more robust representations of simple scenes[10,11], categorize objects[12] and guide our attention[13]. Such tasks may also lead to critical insights into the limits of conscious perception. For example, preserved ensemble information in the relative absence of information about individuals, is thought to show that 'phenomenal' consciousness overflows conscious access[14,15].

Given the fundamental role of ensemble processing, there is immense value in developing process-based models that explain the mechanisms of ensemble extraction. However, so far, mainstream theories of ensemble extraction are largely grounded in verbal descriptions[9]. A known limitation of verbal theories is that they may lack precision of mathematical models, which is requisite for delineating hypothetical constructs and adjudicating between competing theoretical accounts[16,17,18,19].

The goal of the current work is to attempt to fill this gap. We present a theoretical framework and quantitative models of ensemble memory, and compare these models in different experiments to test core process-based hypotheses of how ensembles are computed and represented. We report consistent evidence for a Perceptual Summation model of ensemble memory. According to this model, stimuli evoke distributed patterns of activity over feature values and ensemble representations reflect the global sum of these activations. We find that this model captures a range of phenomena in the ensemble and gist-memory literature.

A major aspect of our modeling framework is that rather than deriving a “best fit” to ensemble data alone, we instead formally link a model of memory for individual items with ensembles. Accordingly, we use our framework to predict performance in a wide range of ensemble tasks, which differ both in stimuli and presentation format, from tasks that involve processing of individual items. Thus, this modeling involves generalizing across different cognitive tasks rather than simply fitting data of a particular task[20,21,22]. The proposed model also provides a high-precision account of human performance by capturing complete distributions of errors in continuous report tasks. Finally, this framework postulates probabilistic mental representations, making it broadly consistent with contemporary population coding models of perception and cognition[23]. Next we review relevant theoretical work on ensemble perception, placing a special focus on dichotomies that highlight core questions researchers may want a model of ensemble processing to answer.

Existing theories of ensemble perception are foundational for stimulating hypotheses of ensemble perception and memory[9,24]. We use these theories to outline three relevant dichotomies, which highlight core desiderata for a quantitative process-model of ensembles.

The first dichotomy is between views that ensembles processing does versus does not involve operating over representations of individual items[24]. This is a core dichotomy because it speaks to how ensemble representations are computed. It also bears on the extent to which ensembles are possible to compute when item information is unavailable to memory, which is critical for theories about the role of ensembles in consciousness[15,25].

One class of views posit that ensemble processing involves pooling over already-processed representations of individual items[10,26,27,28]. According to this view, people have complete representations of individual items and pool them to compute an ensemble. In contrast, other views suggest that ensemble processing involves automatically extracting an average without first representing each individual on its own[6,25]. Researchers also proposed that ensemble extraction involves dividing the total amount of activation elicited by perceived items by their number, without explicitly representing individual items[29].

Much of the work that seeks to address how representations of individual items relate to ensembles is non-quantitative, which can make the connection between individual and ensemble representations difficult to explain. For example, some researchers report that representations of ensembles are present even when memory for individual items is at chance, implying distinct representations for both[10]. However, these authors also report that pooling noisy information about individual items can predict ensemble data[27]. Thus, the question of how representations of individual items relate to ensembles remains an important puzzle in the ensemble literature.

The second relevant dichotomy is between views that noise accrues during an ‘early’ versus ‘late’ stage of ensemble processing[30]. We make two distinctions between possible types of ‘early’ and ‘late’ noise in ensemble processing. The first is between perceptual and post-perceptual noise, that is, noise that accrues during perception without memory demands, versus noise that accrues during memory based processes, such as active maintenance. The second type of distinction is between is presummarization and postsummarization noise[31], which is noise that accrues before versus after ensemble representations are computed.

We distinguish between these two kinds of early and late noise because it is conceptually possible, and is in fact an assumption of the model we propose, that patterns of activation elicited by individual items are corrupted by perceptual noise –consistent with ‘early’ perceptual noise accrual, but that post-perceptual noise accrues after, rather than before ensembles are computed –consistent with ‘late’ postsummarization noise accrual. This view entails that ensemble computation operates over item representations that are corrupted by perceptual but not post-perceptual noise. This contrasts to some subsampling accounts, according to which only a few items are used to create ensemble representations. Current subsampling models are more aligned with the view that ensembles are computed after post-perceptual noise accrues over representations of individual items, which are then used to compute an ensemble when memory is tested[28]. Broadly, these dichotomies between variants of early and late theories of noise provide insight into the time-course of ensemble extraction.

The last relevant dichotomy is between views that ensemble representations are probabilistic versus point estimates. For instance, some researchers examined the content of ensemble-like representations in a visual search task[32]. These authors reported evidence that people are sensitive to the entire underlying (uniform or Gaussian) distribution of features in the external environment, rather than simply an estimate of the average and variance of those features. This claim is consistent with people storing entire probability distributions over visual features, at least in the kind of implicit tasks used in that work[33]. This probabilistic representation view contrasts with an alternative view that people represent a point estimate of the ensemble, such as an average in feature space of each individual item[10,24]. This dichotomy speaks to the richness of ‘summaries’ computed in ensemble tasks.

To preview, the current model posits that ensemble processing involves pooling over individual item representations. It also postulates that representations of individual items are corrupted by noise at an early perceptual stage, but that post-perceptual noise accrues after ensembles are computed. This aligns with the view that ensemble perception is distinct from simply actively maintaining individual items in working-memory and then summarizing them when ensemble memory is probed. Finally, the model posits probabilistic representations, according to which each individual item and the ensemble is represented as a distribution of activity over the entire feature space. Next, we describe the quantitative framework that serves as the conceptual and mathematical basis for the current ensemble model.

We take a top-down strategy for developing a computational model of ensemble processing[34]. We use an existing quantitative framework of memory and carve out a set of plausible constraints on the algorithms that underlie ensemble perception and memory for them. These constraints are formalized using a set of computational models.

We conceived of the current ensemble model using the Target Confusability Competition (TCC) theory of memory[35]. In a set of more than a dozen experiments, TCC outperformed mainstream models of visual memory both in terms of fit to data, and ability to predict data across distinct visual working- and long-term memory tasks. The TCC model combines two fundamental ideas shared by a broad range of cognitive computational models (Figure 2A), which is that memory-based decisions are made under uncertainty[36,37] and that information in the world is processed based on its psychophysical, rather than physical similarity structure[38,39,40].

The first premise of TCC is that memory representations are intrinsically probabilistic and vary in strength, a core principle of Signal Detection Theory[37,41,42] and, broadly, Bayesian models of cognition[43,44,45,46]. For instance, a remembered item is assumed to be neither completely forgotten nor completely remembered. Instead, there is a probability distribution over how well the item is remembered, such that sometimes it is remembered with high fidelity and elicits a strong familiarity signal, and other times it is remembered with lower fidelity and elicits a weaker familiarity signal. Within the Signal Detection theory and TCC the strength of each memory’s familiarity signal is captured with the signal-to-noise ratio parameter, d'.

The second premise of TCC is that familiarity spreads across feature space according to the stimulus' psychophysical properties, an assumption shared with other foundational models of memory[47,48,49,50,51]. Specifically, the familiarity of a given stimulus is a function of the psychophysical similarity between this stimulus and contents of memory, which can also be thought of as distributed patterns of activation in neural populations that are selective to remembered features values[52,53]. This assumption entails that stimuli will elicit a stronger familiarity signal if they are more psychophysically similar to contents in memory. For instance, if the remembered item is a purple square, the color purple will elicit a very strong familiarity signal, as will colors that are nearly perceptually indistinguishable from purple. Colors that are somewhat similar to purple, such as magenta, will also elicit a familiarity signal, which will be stronger than those elicited by relatively dissimilar colors, such as green. This latent psychophysical similarity function and the corresponding distribution of memory signals is approximately exponential in form, in line with previous theories of memory and generalization[39,40,54,55].

To summarize, TCC is a model that formally combines two fundamental views about memory processes in a way that permits generalization across memory tasks with a single free parameter, d'. The generalizability and parsimony of TCC, as well as its basis on probabilistic models of cognition and psychophysical scaling, make it a powerful framework for building cognitive architectures. We use TCC to derive the Perceptual Summation model as well as a set of contending models, with which we test hypotheses of how ensembles are computed and represented. We also derive and test a set of alternative, non-TCC based models that make different processing assumptions. Next, we describe the TCC working-memory model for individual items, how we extend it to models of ensemble memory, and formally link these models.

Figure 2B (left) shows a schematic of a typical trial in a visual working-memory task for individual items, which requires memorizing three colored circles and their spatial locations. The TCC model postulates that each item elicits some location dependent pattern of activity, which causes an increase in familiarity for its respective color, but also for similar colors. These levels of activation are each corrupted by perceptual noise, which makes it more difficult to distinguish highly-similar feature values from one another. Throughout our modeling, we assume that perceptual noise affects individual item activations in the same way in both working-memory and ensemble tasks (see Methods for how perceptual noise was measured and modeled).

After individual item representations are perceived, their pattern of activity is corrupted by attention- and memory-based noise. These effects of post-perceptual noise are captured with a single free parameter d', which quantifies the signal-to-noise ratio of each individual representation. The signal-to-noise ratio is affected by key experimental variables, such as memory load, encoding time, and the retention interval, each of which affects how well the items are initially encoded and how much noise accumulates during memory maintenance. At the end of the trial, the probed-item's location is queried, and participants report on the color channel that generates the maximum familiarity signal.

Formally, the TCC model for individual items is given by the following equation:

$$r_{i,VWM} = argmax\Big(f(x)_i d' \quad + \sigma_{Noise}\Big). \qquad (1)$$

The index i denotes, the probed item, $r_{i,VWM}$ is the predicted response on the continuous report visual working memory task for that item, $f(x)$ is the measured similarity of each color x with respect to item i, d' is a free parameter that quantifies the signal-to-noise ratio, σ_{Noise} is a fixed amount of post-perceptual noise (set to one standard deviation with no loss in generality), and $argmax$ denotes the decision rule that memory reports are based on the feature that generates the maximum familiarity signal.

We developed the Perceptual Summation ensemble model from the TCC model for individual items, as well as constraints based on prior evidence from the ensemble literature. These constraints include seemingly contradictory evidence that memory for individual items can predict memory for ensembles, but that memory for ensembles is more robust than memory for individual items[10].

Like the model for individual items, the Perceptual Summation model postulates that each item in the memory array elicits patterns of activity over feature values, each of which is corrupted by perceptual noise (Figure 2B, right). However, the Perceptual Summation model postulates that the ensemble is extracted during encoding, before memory-based noise accrues over representations of individual items. Thus, the model postulates that memory-based noise accumulates over the ensemble instead of representations of each item in the array. When probed on the average, participants report on the color channel that generates the maximum familiarity signal. The equation for the Perceptual Summation model is the following:

$$r_{ENS} = argmax\left(\ (\sum_{i=1}^{N} f(x)_i \ \ d' \ ) \ + \sigma_{Noise} \right). \qquad (2)$$

Note that Equations 1 and 2 are nearly identical, with the exception that self-reports on the visual working memory task are determined by levels of activation elicited by a single probed item (i), whereas self-reports on the ensemble task (r_{ENS}) are determined by the summed levels of activation of all N items.

With this framework we connect the model for individual items and ensembles. We postulate that the patterns of activation elicited by each item in the memory array is the same in both working-memory and ensemble tasks, and is pooled via summation in the early perceptual stage of ensemble extraction. Like in the model for individual items, this pattern of activation is measured with a psychophysical similarity function, which captures how familiarity signals are distributed across feature values for each item, and a single free parameter d', which measures the signal-to-noise ratio that scales these patterns of activation based on the demands of the memory task. We formally link memory for individual items and ensembles by estimating the signal-to-noise ratio (d') of each individual item from a visual working-memory task for individual items, and substituting this signal-to-noise ratio into the Perceptual Summation model to compute the predicted summed pattern of activation of the ensemble. With this approach, we predict entire distributions of memory errors in continuous report ensemble tasks with zero free parameters.

To summarize, the difference between representations of individual items and ensembles, is that in ensemble tasks patterns of activation elicited by individual items are pooled via summation before post-perceptual noise accrual. This entails that the signal-to-noise ratio of the post-summation ensemble representation will be larger than it is for individual items when there is overlap in feature values, or redundancies between items in the ensemble array. Through the lens of likelihood signal detection theory[41], this pooling mechanism can be seen as an optimal way of combining the likelihood elicited by each item into a more robust ensemble memory representation, or gist, as opposed to treating the evidence elicited by each item separately.

The Perceptual Summation model's pooling mechanism can be seen as a cognitive-level approximation of processes described in neural population coding models. Very generally, some evidence suggests that increased population size may increase the amount of information embedded in populations of neurons[56]. Although the relationship between population size and readout accuracy is extremely complex and an active topic of investigation[56,57,58,59], this framework provides one neurally plausible instantiation for the computations postulated in the Perceptual Summation model.

To test the predictions of the Perceptual Summation model we compare it to a set of alternative models. The first prediction we consider is the time-course of ensemble extraction. The Perceptual Summation model's 'early pooling' prediction contrasts with an alternative view that individual items are held in working-memory until ensemble memory is probed, at which point they are pooled to compute an ensemble. This alternative view is informally embodied in some subsampling theories of ensemble processing[28]. We formalize this prediction within the TCC framework with the Post-perceptual Summation.

The Post-perceptual Summation model predicts that people maintain location dependent representations of each item in memory – as they would in a standard working-memory task for individual items – until they are probed on their memory for the ensemble. Thus, according to this model, ensemble representations are computed at a relatively late stage, and therefore, each item across memory-based noise separately, before the ensemble is pooled. The equation for the Post-perceptual Summation is the following:

$$r_{ENS} = argmax\left(\sum_{i=1}^{N} (f(x)_i \;\; d' \; + \; \sigma_{Noise}) \right). \qquad (3)$$

Note that the terms in Equations of the Post-Perceptual (3) and Perceptual Summation (2) models are nearly identical, with the difference that summation occurs over individual items that already accrued post-perceptual noise (3), versus before representations of individual items have accrued post-perceptual noise (2). To summarize, these two models can mimic each other if d' is allowed to freely vary, however, because we use a generalization approach, d' is constrained across tasks, allowing us to differentiate these models (see Supplementary Information for extended discussion of these issues).

Finally, we consider an ensemble model that follows from theories that ensemble averages are extracted automatically, without processing of individual items[6,25], which we refer to as the Automatic Averaging model. Although still nested within the TCC framework, this model differs from the Perceptual and Post-perceptual Summation models because it postulates that individual items in ensemble tasks automatically elicit distributed patterns of activation around the average feature value in ensemble array, rather than eliciting item-specific patterns of activation that are pooled via summation. This representation of the average is also probabilistic and scaled by the signal-to-noise ratio of a single memory representation. We also consider alternative assumptions about the signal-to-noise ratio for this model in the Supplementary Material. To summarize, this model is equivalent to assuming that the ‘average’ is directly perceptually available to people in the same way that an item that is physically present is. As shown in Equation 4, this model postulates that people extract a single probability distribution over the mean feature, which is also corrupted by noise,

$$r_{ENS} = argmax\left(f(x)_{Mean}\, d' \; + \; \sigma_{Noise}\right) \qquad (4)$$

Note that Equation 4 is nearly identical to Equation 1 with the exception that the similarity function is centered on the average feature value, instead of the value of an individual item. Figure 3 depicts each of these TCC ensemble models.

So far, these ensemble models posit that, on average, each item is weighted equally when computing an ensemble. This is tenable under conditions that do not lead to disproportionate prioritization of a specific item, or subset of items[60,61].

However, it is known that some conditions do elicit unequal weighting of items in memory. For instance, items that are shown more recently tend to be remembered better than items shown less recently, and such recency effects affect ensemble representations as well[62]. To evaluate the generalizability of our modeling we extended it to conditions in which items receive unequal prioritization in memory. Furthermore, the summation account becomes more distinct from other possible accounts when items vary in strength, therefore, this analysis also provides a stronger test for the view that ensemble representations reflect a sum of local patterns of activation. Finally, this analysis helps demonstrate that we can predict both inter- and intra-individual variations in ensemble processing.

To this end, in one of our experiments we used a sequential presentation ensemble paradigm. One way to generalize the TCC-based models to this situation is to simply obtain separate d' estimates for each item in the sequence and use these estimates to compute ensemble predictions. However, we can also use a temporal model that captures memory changes as a function of the sequential presentation with fewer parameters. We used prior modeling work[62] to extend our modeling in this way (see Methods). As expected, we found the same pattern of results using both types of models. Next, we describe a few alternative, non-TCC models of ensemble perception.

Currently there are no computational models of ensemble processing that fully capture distributions of errors in a continuous self-report task and that can account for data across a range of ensemble manipulations. However, in order to bolster the interpretability of our modeling, we derived a set of alternative models that serve as conceptual foils to the TCC ensemble models. Some of these models are baseline models that make extremely simplistic assumptions about ensemble processing, which we use to check the tenability of our TCC models. Other models link memory performance for individual items to memory for ensembles while postulating different assumptions about memory processes, such as that there are true 'guessing states'[64]. Each of these models is depicted schematically in Figure 4. Because we do not find that these are best-performing models, for ease of exposition, we include a conceptual description of these models in the Methods.

We ran five experiments to evaluate the predictive accuracy of each ensemble model, with the goal of assessing the generalizability of our modeling results across different ensemble tasks. As previewed, each experiment had the same structure, meaning that participants completed one block of a visual working memory task and one block of an ensemble task (presented in random order across participants). This allowed us to measure d' in the visual working-memory task, and to use it to predict performance in the ensemble task. In Experiments 1 and 2 we examined people's memory for color (Figure 5A), and manipulated set size and the range of colors values in the ensemble task, respectively. In Experiments 3 and 4 we evaluated the generalizability of these results for a higher-level shape feature space (Figure 5B), where we also manipulated set size and varied the range of shapes, respectively. Finally, in Experiment 5 we used a sequential presentation task (Figure 5C), to test models when memory representations receive different priority.

Results

Our goal is to evaluate the ability of the Perceptual Summation ensemble model to generalize performance from the visual working-memory task to the ensemble task. We formally compare the predictive accuracy of this model with other models using the predicted negative log likelihood (PNLL) between it and the contending models. PNLL is a predictive model comparison metric because we assess the models based on their capacity to generalize across tasks, that is, make zero-free-parameter predictions on new data in a different task. However, we do note that all TCC models also yield 'good fit' to data ($R^2 \geq .9$ across all experiments; see Supplement). Because PNLL is a negative log likelihood, lower scores reflect less deviance and better model predictions. PNLL naturally accounts for model complexity because it captures predictive accuracy rather than goodness of fit (for an elaborated discussion of this point see Supplement).

For our main analysis with TCC ensemble models, we fit the TCC visual working-memory model to data and substituted d' estimates from these fits into the ensemble models to predict the ensemble data. To ensure the robustness of our models' performance, we also implemented a reverse inference analysis in which we fit the ensemble models to the ensemble data and then use best fitting parameters from the ensemble task to predict the working-memory data (Supplement). We find that our results are robust across these different methods of prediction.

We implement analyses at the level of individual participants. Specifically, we compare the observed PNLL between each model and the best performing model using a paired t-test. Data distribution was assumed to be normal but this was not formally tested. We report the observed effect size (d_z) and the lower and upper bounds of a confidence interval for the mean difference (CI_L and CI_U, respectively). We use a conservative Bonferroni correction[64] to control for multiple comparisons. For our main comparisons in Experiments 1-4 and Experiment 5 there are six and eight family-wise comparisons (m), respectively, and the adjusted significance threshold ($\alpha_A = \alpha/m = .05/m$) is .008 and .006, respectively. For our reverse inference comparisons there are two family-wise comparisons in each experiment and $\alpha_A = .025$. We find that each central comparison was statistically significant when adjusting for multiple comparisons.

Ensemble memory for color with different set sizes. In Experiment 1, participants completed a visual working-memory and ensemble task using color as the stimuli, and manipulated set size to assess how each model captures changes as a function of memory load. We found that the Perceptual Summation was the best-performing model.   Figure 6 shows the difference in PNLL between the Perceptual Summation and competing models, and fit statistics. Figure 7 shows fits of TCC models to aggregate and example individual data.

Ensemble memory for color with different ranges. In Experiment 2, participants performed a color task and we manipulated the range of the colors in the ensemble task, that is, how distinct they were from each other. We found that the Perceptual Summation model was the best-performing model. Figure 8 shows the difference in PNLL between the Perceptual Summation and competing models, and fit statistics. Extended Figure 1 shows fits of TCC models to aggregate and individual data.

Ensemble memory for shapes with different set sizes. In Experiment 3, we manipulated set size and had participants remember shapes instead of colors. We found that the Perceptual Summation was the best-performing model. Extended Figure 2 shows the difference in PNLL between the Perceptual Summation and competing models, and fit statistics. Extended Figure 3 shows fits of TCC models to aggregate and individual data.

Ensemble memory for shapes with different ranges. In Experiment 4, we manipulated the range of shapes in the ensemble task. We found that the Perceptual Summation was the best-performing model. Extended Figure 4 shows the difference in PNLL between the Perceptual Summation and competing models, and fit statistics. Extended Figure 5 shows fits of TCC models to aggregate and individual data.

Ensemble memory for sequentially presented stimuli. In Experiment 5, we presented stimuli sequentially, introducing variation in the strength of the items. We found that the Recency Perceptual Summation was the best-performing model. As expected, this model performed comparably to a model where we measured a separate d' for each item in the sequence. Figure Extended Figure 6 shows the difference in PNLL and statistical comparisons. Extended Figure 7 shows fits of TCC models to aggregate and individual data.

General Discussion

Across five experiments we find support for a Perceptual Summation ensemble model that postulates that ensemble representations are a sum of activations elicited by individual items in the memory array, which are pooled at a relatively early, encoding stage of processing. We use the TCC framework to formally link a working-memory model for individual items with this ensemble model. The Perceptual Summation model yields zero-free-parameter predictions of the full distribution of errors in ensemble tasks, using parameters obtained from a matched visual working memory task for individual items. It is a general process-model of ensembles, developed based on an existing theory of memory for individual items to make predictions for any ensemble task. Our modeling demonstrates that it can make predictions for ensemble tasks that use different stimuli spaces and presentation formats. In the Supplement we report simulations that demonstrate how the model can be extended to other tasks.

We compared our Perceptual Summation model of ensembles to a suite of contending models to adjudicate between competing hypotheses regarding how ensembles are extracted. The first critical comparison is between the Perceptual and Post-perceptual Summation models, which provides insight into the time-course of ensemble extraction. The Perceptual-Summation model entails that people pool over individual item representations relatively early at the perceptual/encoding stage of ensemble extraction. In contrast, the Post-perceptual Summation model entails that people pool at a later processing stage, after individual items are encoded and consolidated in working-memory[28]. We found that the Perceptual outperformed the Post-perceptual Summation model, indicating ensemble processing in these settings is more akin to a perceptual process, rather than a complex deliberate process, in which people calculate a pooled representation using individual memory representations when their ensemble memory is probed.

Second, we found that the Perceptual Summation outperformed the Automatic Averaging model. The Automatic Averaging model aligns with prior proposals[6] that people extract an average without maintaining representations of individual items; it serves as a logical foil to the Perceptual and Post-perceptual Summation models, which both predict that ensembles are constructed from representations of individual items. Across all studies, we found that the Perceptual Summation model outperformed the Automatic Averaging model, suggesting that people use representations of individual items to extract ensembles, rather than automatically extracting an average.

We also compared the Perceptual Summation model with four non-TCC models, which elucidates how to characterize ensemble representations. We find that the fully probabilistic TCC models outperform point estimate and partial distribution models. We clarify here that when we refer to representations as ‘probabilistic’ we do not assume that they must conform to classic probability axioms[65]. Rather we assume that memory representations preserve uncertainty information for the full distribution of feature values, and that performance in memory tasks reflects a readout of these uncertainties over feature values. These results are broadly consistent with neural population coding models of memory, according to which memory representations are grounded in distributed neural patterns of activation across feature values[53,66,67].

We conclude this section by noting that, like in all model comparisons, our inferences are qualified by the set of models we consider. For instance, we do not make the strong claim that there are no alternative Automatic Averaging, or Point Estimate models that could provide a better account of the data in principle. Our goal is to develop a broad range of alternative models within and outside of the TCC framework, with varying assumptions, and implement them as fairly to each theoretical position as possible. In the Supplement we discuss how our model connects and differs to existing models of ensemble processing. We anticipate that future modeling work may provide a new suite of alternatives. We believe it is critical that such work focuses on developing models that can account for performance across a range of ensemble tasks, have the potential to generalize across task structures and make high precision predictions of performance.

The Perceptual Summation model has relevance for theories of gist memory. Gist memory is broadly defined as memory for ‘generalities’ across multiple items, as opposed to memory for individual items and, as such, both gist and ensemble processes involve abstracting regularities from multiple items[24,68]. Our model of ensembles cannot speak to how memory for gist and individual items interact during short or long-term memory retrieval. However, it provides a candidate explanation for how gist memory representations are computed. According to this model, the bottleneck during encoding of individual items is the same across visual working-memory tasks for individual items, and the extraction of a pooled representation. Furthermore, the model proposes that memory-based noise accrues in the same way regardless of whether people are instructed to remember a single item or ensemble. Critically, it postulates that pooling of representations occurs at a relatively early processing stage, prior to post-perceptual noise accrual. Together, the Perceptual Summation provides a parsimonious and precise account of how gist representations may arise from representations of individual items, while still being more robust than representations of individual items.

This model also provides an unambiguous account regarding how processing of individual items differs from processing of gist[69]. That is, instead of using theoretically underspecified constructs[70], such as focused versus diffuse attention[30,71], or preattentive and attentive modes of processing[72,73], it describes how different computations over the same representations can give rise to distinct types of memories. We believe such an approach has great promise for building precise and testable models of gist memory, hierarchical representations and reconstructive memory processes in the visual domain.

We conclude the article by discussing a few potential limitations and venues for future research. In the current modeling approach we use a single parameter, d' to measure a potentially diverse set of processes. In line with standard Signal Detection models, d' quantifies the signal-to-noise ratio of each memory representation, and different processes at encoding, maintenance and retrieval are built into this measure. However, signal detection-based accounts are fully compatible with the view that processes and memory representations are multi-dimensional[41]. Our measure of memory with d' simply captures how people combine multidimensional processes and memory representations into a single decision variable, which they use to make memory judgments when their memory is probed[74]. Naturally, this measure can be complemented with other modeling frameworks that unpack these processes. We elaborate and clarify related aspects of Signal Detection Theory, TCC, and our generalization approach in the Supplement.

Another limitation is that we did not model all possible phenomena in the ensemble literature. This is because our goal was to formally establish a link between two different processing models and, to this end, we focused on a set of mainstream ensemble tasks where the patterns of effects are robust. In the Supplement, we report simulations to demonstrate that our model can, in principle, capture both outlier discounting or increased weighting of outliers. We also report simulations to show how the model accounts for differential effects of set size on the fidelity of ensemble representations, and effects of various distributions of stimuli in the ensemble array. Together, our aim is to lay out a theoretical and methodological framework for future modeling research of ensemble and gist memory.

Methods

The study was completed online through the University's SONA system and approved by the Institutional Review Board. All participants were at least 18-years old, provided informed consent and reported normal or corrected-to-normal vision. Participants were from the University of California, San Diego community and participated in exchange for course credit. Participants were blind to the hypotheses of the study. In each experiment, we collected data until our final sample size was *n*=50, which affords 99% power for a medium effect size ($d_z = .5$) for a paired *t*-test at $\alpha = .05$. We did not analyze data of participants who failed to complete the study. We also excluded data from participants if their d' estimates in any of the visual working memory task conditions (and for the last item in the sequential presentation task) were below 1.5 standard deviations of the group mean. All data and are available in the Open Science Framework repository (https://osf.io/vx6pc/) and code will be made publicly available upon publication.

Experiment 1: Memory for color with manipulation of set size Participants completed a block of a visual-working memory task and an ensemble task (order of blocks was randomized across participants). At the beginning of every trial in both the visual-working memory and ensemble tasks, participants were shown a written prompt with the current trial number and the total number of trials in that block (1,000ms). After the prompt offset, participants were shown a fixation cross in the center of the screen and (six) placeholders (1,000ms). Next, participants were briefly presented with the memory array (350ms). We manipulated memory load (randomly across trials) in the visual working memory and ensemble tasks, thus, participants were instructed to remember six (50% of trials) or eight items in both tasks. The color of each circle was randomly sampled from the CIELAB color space of Schurgin et al. (2020) with the constraint that each color had to be at least 30˚ away from other colors in the array. In the ensemble block, the memory array also consisted either of six (50% of trials) or eight colored circles. The step-size between colors in the ensemble task was fixed to 15˚ for both set sizes. There were 150 trials in the visual working memory block and 150 trials in the ensemble block (75 trials per memory load condition in each of the tasks).

The memory array in both blocks was followed by a retention interval (900ms) and the memory probe. In the visual-working memory task, participants were shown a black outline around one of the placeholders, which cued them to report on the color of the circle shown in that spatial location. In the ensemble task, participants were instructed to report on the average color. In both tasks, participants reported on the color using a color wheel.

Experiment 2: Memory for color with manipulation of range in the ensemble task The procedure of Experiment 2 was identical to the procedure in Experiment 1 with the following exceptions. First, in both the visual-working and ensemble memory block, the memory array always consisted of six colored circles. Second, in the ensemble task, the stepsize between colors in each condition was constrained to be 10˚ (60˚ range condition), 15˚ (90˚ range condition), or 20˚ (120˚ range condition). There were 75 trials of the visual-working memory task, and 225 trials of the ensemble task (75 trials in each of the three range conditions).

Experiment 3: Memory for shape with manipulation of set size Experiment 3 was identical to Experiment 1, with the exception that people were shown shapes instead of colors and we changed encoding time to 1,000ms and the retention interval to 800ms because the shape task is more difficult than the color task. Shape stimuli were taken from Li, Liang, Lee & Barense (2020).

Experiment 4: Memory for shape with manipulation of range in the ensemble task Experiment 4 was also identical to Experiment 2, with the exception that people were shown shapes instead of colors and we changed encoding time to 1,000ms and the retention interval to 800ms.

Experiment 5: Memory for sequentially presented colors The goal of Experiment 5 was to model data from a sequential instead of simultaneous presentation paradigm. Therefore, in this experiment participants were instructed to remember colors of colored pictures of real-world objects[43]. We used pictures of real-world objects instead of uniform stimuli (e.g., circles) because this allowed us to easily probe an item's serial position in the sequential visual-working task for individual items by showing participants a grayscale photo of one of the objects in the sequence and probing them on that object's color.

As before, all participants completed a block of the visual working memory and ensemble task. In both tasks, participants self-advanced each trial by mouse-clicking on a fixation cross in the center of the screen. The mouse-click was followed by a brief delay (1,000ms), after which they were shown a sequence of six objects, each presented one at-at-time in the center of the computer screen. In both tasks, each object was presented for 600ms and followed by a 450ms inter-stimulus interval. In the visual-working memory task, on each trial, each object in the sequence was unique, and the color of each object was constrained to be at least 30° away from colors of the remaining object. In the ensemble task, on each trial, each object in the sequence was the same (though different objects were presented across trials), and the stepsize between colors was 20°. In order to measure effects of recency in the ensemble task, we adapted a manipulation from prior work[62]. Specifically, on half of the trials the first (or last) three objects in the sequence had colors that were counterclockwise to the mean color, whereas the last (or first) three objects had colors that were clockwise from the mean color in color space.

In both tasks, the last object in the sequence was followed by a 900ms delay. Within the delay period participants were shown a dynamic visual mask, which was displayed for 100ms, 100ms after the last object offset. The mask was used to reduce potential effects of iconic memory on recency effects in the sequential presentation design. After the retention interval, participants were probed on their memory with a continuous report. In the visual-working memory task, participants were shown a grayscale version of one of the six objects in the sequence and instructed to adjust its color to its color on that trial. In the ensemble task, participants were shown a grayscale version of the object from that trial and instructed to adjust it to have average color on that trial. There were 120 trials in the visual working memory task and each object in the sequence was probed equiprobably (on 20 trials) across the experimental block. There were 96 trials in the ensemble task, with 48 trials in the counterclockwise and clockwise conditions.

Generating predictions from TCC ensemble models. Models were fit separately to each participant's visual working memory data. The best-fitting parameter estimates from these fits were used to predict the same person's data on the ensemble task. In Experiments 1-4, we fit the standard TCC model for single items to the visual working memory data. The formula for this model is given in Equation 1. After obtaining a d' estimate from fitting models to the visual-working memory data, we substituted this parameter into Equations 2-4 of the Perceptual and Post-perceptual Summation and Automatic Averaging model, to predict the ensemble data. In Experiment 5, we fit the sequential version of the visual-working memory model (see below: Equation 5), and substituted both the d' and rate parameters into Equations 6 and 7 to predict the ensemble data using the Recency variants of the Perceptual and Post-perceptual Summation models for ensembles, respectively. As noted, the Automatic Averaging model postulates that people extract a single representation of the mean without building it up from representations of individual items, therefore, for this model we only used a single d'estimate to make predictions.

Generating predictions from Non-TCC ensemble models. The first alternative non-TCC ensemble model is the Noise-Free Point Estimate model, according to which people automatically extract a *noise-free* point estimate of the mean feature, which is corrupted by motor noise only. In other words, the Noise-Free Point Estimate model simply predicts that self-reports on the ensemble task are the true mean ($\overline{X}$). This model is unlikely to perform well since it cannot capture the full distribution of errors in a delayed estimation task, however, we include it because it serves as a logical reference point against which to compare the assumptions of TCC ensemble models, such as that ensemble representations are probabilistic.

To generate predictions from the Noise-free point estimate model we calculated an equally weighted average value of the ensemble (like in the Automatic Averaging model). We then added a small amount of jitter to this estimate to simulate small effects of motor noise. To simulate motor noise, we used the built-in `randn` function in MATLAB, which generates random samples from a standard Gaussian distribution.

The second non-TCC model is the Noisy Point Estimate model, a more plausible extension of the Noise-Free Point Estimate model. According to this model, people automatically extract a point estimate of the mean feature along with a uncertainty interval around this value. More precisely, we make the simplifying assumption that people represent a fixed uniform uncertainty interval around the true mean value[65]. This model provides a simple way of capturing the idea that people represent a noisy representation of the ensemble. To generate predictions from the Noisy Point estimate model we drew random samples of data from a uniform distribution, which had a range of 60˚ and was centered on the true value of the average i.e., samples were drawn from -30 and 30˚ around the mean value. The assumption behind this model is that people represent a uniform uncertainty interval around the true mean value, thus, the number of samples was based on the number of trials in each ensemble condition.

So far, neither of the first two Point estimate models formally link memory for individual items to memory for ensembles. Therefore, we consider two additional models, which do link processing across the working memory and ensemble task, and make more tenable assumptions about processing on ensemble tasks. The first of these models we refer to as the Average Item Point

Estimate model, according to which people compute ensembles by averaging over point estimates of individual item representations, a model closely related to the averaging view[6,10,24].

The second of these models is the Precision Ensemble model, according to which people maintain a point estimate of the mean, which has a Gaussian, rather than uniform uncertainty interval around it. Rather than using a fixed interval for all subjects as we did for the Noisy Point Estimate model, we calibrate this interval for each individual by obtaining a standard deviation estimate from their working memory data using the popular standard mixture model (inspired by Zhang and Luck, 2008). Therefore, this model also inherits a fundamentally different processing assumption about memory than the TCC ensemble models, which is that there are true 'guessing states' in memory, such that there is no evidence that can be used to report on memory. To generate predictions for the Average Item Point Estimate we used people's working memory data to sample n point estimates (measured in degrees of error) for each item in the array, and then averaged across these. For instance, if people had to remember six items on the visual working memory and ensemble tasks, we drew (with replacement) six samples of their self-report data (converted to error in degrees) on the visual working memory task, and averaged across these. This was repeated for each trial, to generate the predicted distribution of errors on the ensemble task.

To summarize, neither the Noise-free Point Estimate nor the Item Average Point Estimate model postulate probabilistic representations. The Noisy Estimate and Precision ensemble model postulates partially probabilistic representations because there is an uncertainty interval around the mean, but not a full probability distribution over feature values[65]. Note that we do not presume that these point-estimate models capture all possible ways in which point-estimate or partially probabilistic models could account for the data. However, in the absence of other quantitative models in the literature that can apply to such ensemble perception tasks, we created models that spanned a wide range of plausible assumptions regarding the nature of ensemble extraction and the properties of ensemble representations. We revisit this point in the General Discussion.

The Precision Ensemble model was implemented by fitting a Standard Mixture model[63] to each individual's visual working memory data and using the standard deviation estimate from this model to compute an uncertainty interval around the true average. The latter was implemented by sampling random samples of data (based on the number of trials on the ensemble task), from a normal distribution with mean zero and the standard deviation set to the standard deviation (inverse of precision) estimate from the mixture model.

Recency TCC ensemble models. The Recency TCC model quantifies recency weights using as an exponential function (without base e) over the serial position of each stimulus in the sequence. The recency model for individual items is given by the following equation:

$$r_{i,VWM} = argmax\left(f(x)_i \, d' \; rate^j \; + \sigma_{Noise} \right), \tag{5}$$

where all terms are identical to those given in Equation 1, with the exception of a second parameter $rate$, which has the item position of item $j$in the exponent (where j=1 is the most recent item in the sequence). The rate parameter is a free parameter bounded between 0 and 1 that captures the effects of memory decay on memory, with smaller values (of the parameter) indicating stronger decay effects and, therefore, relatively higher weighting of more recent items in the ensemble (and relatively better performance for them in the visual working memory task). The equations for the Perceptual and Post-perceptual Summation ensemble models are extended in a similar way, as shown below (Equations 6 and 7, respectively):

$$r_{ENS} = argmax\left(\left(\sum_{i=1}^{N} f(x)_i \;\; d' \;\; rate^j \right) \; + \sigma_{Noise} \right), \tag{6}$$

$$r_{ENS} = argmax\left(\left(\sum_{i=1}^{N} f(x)_i \;\; d' \;\; rate^j \;\; + \sigma_{Noise} \right) \right). \tag{7}$$

Note that the Automatic Averaging model is unchanged because it postulates that people automatically extract a single representation of the mean without building up this representation from individual items. We treat this model as being conceptually equivalent to a prototype model that entails an equally weighted average[75].

Model fitting. All models were fit to visual-working memory data in MATLAB using Maximum Likelihood Estimation (MLE) by minimizing the negative log likelihood. Minimization was implemented with the `fmincon` algorithm in the Optimization Toolbox as well as basic iterative search. The predictive accuracy of each model was measured using the predictive negative log likelihood (PNLL). For the main analyses, we substituted best-fitting parameters from the visual-working memory task into equations for each of the ensemble models to predict data on the ensemble task, and calculated the PNLL using data from the ensemble task conditions.

Psychophysical similarity function and perceptual noise. The psychophysical similarity functions in our color experiments were estimated in prior work using a Likert task and verified using a ‘triad’ task, a mainstream method for obtaining psychophysically scaled similarity data[35].Perceptual noise was measured with a perceptual matching task, also a mainstream task for quantifying perceptual confusability of visual stimuli. In this task participants were shown a color, and asked to match it to one of 60 colors (6 degrees apart) presented simultaneously on the computer screen. This task provides insight into how perceptual noise affects the perceptual confusability of stimuli. These perceptual matching data were converted to a covariance matrix, which was convolved with the psychophysical similarity function[35]. We note that the resulting psychophysical similarity Signal Detection model is a simulation based approximation of a correlated noise signal detection model[76].

The same tasks were applied to the shape data. The shape wheel was created and validated as a circular space in prior work[77]. We collected Likert similarity data and perceptual confusion data for this wheel using the same methods as used for the color data.

Data availability statement

Data is publicly available at the following OSF link:
https://osf.io/mt29p/?view_only=5b3035a70b194e9a8a8a926ec639fad9.

Code Availability Statement

Code is publicly available at the following OSF link:

References

1. Baddeley, A. (1992). Working memory. *Science*, 255(5044), 556–559.

2. Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, *63*(2), 81.

3. Pashler, H. (1984). Processing stages in overlapping tasks: Evidence for a central bottleneck. *Journal of Experimental Psychology: Human Perception and Performance*, 10(3), 358–377. https://doi.org/10.1037/0096-1523.10.3.358

4. Simon, H. A. (1990). Invariants of human behavior. *Annual review of psychology*, *41*(1),1-20.

5. Kahneman, D. (2003). A Psychological Perspective on Economics. *American Economic Review*, 93(2), 162–168. https://doi.org/10.1257/000282803321946985

6. Ariely, D. (2001). Seeing Sets: Representation by Statistical Properties. *Psychological Science*, 12(2), 157–162. https://doi.org/10.1111/1467-9280.00327

7. Brady, T. F., & Oliva, A. (2008). Statistical Learning Using Real-World Scenes. Psychological Science, 19(7), 678–685. https://doi.org/10.1111/j.1467-9280.2008.02142.x

8. Goldstein, M. H., Waterfall, H. R., Lotem, A., Halpern, J. Y., Schwade, J. A., Onnis, L., & Edelman, S. (2010). General cognitive principles for learning structure in time and space. *Trends in Cognitive Sciences*, 14 (6):249-258.

9. Whitney, D., & Yamanashi Leib, A. (2018). Ensemble Perception. *Annual Review of Psychology,* 69(1), 105–129. https://doi.org/10.1146/annurev-psych-010416-044232

10. Alvarez, G. A., & Oliva, A. (2008). The Representation of Simple Ensemble Visual Features Outside the Focus of Attention. *Psychological Science*, 19(4), 392–398. https://doi.org/10.1111/j.1467-9280.2008.02098.x

11. Brady, T. F., Shafer-Skelton, A., & Alvarez, G. A. (2017). Global ensemble texture representations are critical to rapid scene perception. *Journal of Experimental*

Psychology: Human Perception and Performance, 43(6), 1160–1176. https://doi.org/10.1037/xhp0000399

12. Utochkin, I. (2015). Ensemble summary statistics as a basis for visual categorization. *Journal of Vision*, 15(12), 891. https://doi.org/10.1167/15.12.891

13. Balas, B., Nakano, L., & Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision*, 9(12), 13. https://doi.org/10.1167/9.12.13

14. Block, N. (2011). Perceptual consciousness overflows cognitive access. *Trends in cognitive sciences*, *15*(12), 567-575.

15. Cohen, M. A., Dennett, D. C., & Kanwisher, N. (2016). What is the Bandwidth of Perceptual Experience? *Trends in Cognitive Sciences*, 20(5), 324–335. https://doi.org/10.1016/j.tics.2016.03.006

16. Grahek, I., Schaller, M., & Tackett, J. L. (2021). Anatomy of a psychological theory: Integrating construct-validation and computational-modeling methods to advance theorizing. *Perspectives on Psychological Science*, 1745691620966794.

17. Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, 16(4), 789–802. https://doi.org/10.1177/1745691620970585

18. Navarro, D. J. (2021). If mathematical psychology did not exist we might need to invent it: A comment on theory building in psychology. *Perspectives on Psychological Science*.

19. Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, 26(5), 1596–1618. https://doi.org/10.3758/s13423-019-01645-2

20. Busemeyer, J. R., & Wang, Y. M. (2000). Model Comparisons and Model Selections Based on Generalization Criterion Methodology. *Journal of Mathematical Psychology,* 44(1), 171–189. https://doi.org/10.1006/jmps.1999.1282

21. Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, 55, 1–7.

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

23. Rust, N. C. (2014). Population-Based Representations. *The cognitive neurosciences*, 337.

24. Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, 15(3), 122–131. https://doi.org/10.1016/j.tics.2011.01.003

25. Ward, E. J., Bear, A., & Scholl, B. J. (2016). Can you perceive ensembles without perceiving individuals?: The role of statistical perception in determining whether awareness overflows access. *Cognition*, 152, 78–86. https://doi.org/10.1016/j.cognition.2016.01.010

26. Oriet, C., Giesinger, C., & Stewart, K. M. (2020). Can change detection succeed when change localization fails? *Journal of Experimental Psychology: Human Perception and Performance*, 46(10), 1127–1147. https://doi.org/10.1037/xhp0000834

27. 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

28. 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

29. ŠEtić, M., ŠVegar, D., & Domijan, D. (2007). Modelling the statistical processing of visual information. *Neurocomputing*, 70(10–12), 1808–1812. https://doi.org/10.1016/j.neucom.2006.10.069

30. 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

31. Solomon, J. A. (2021). Five dichotomies in the psychophysics of ensemble perception. *Attention, Perception, & Psychophysics*, *83*(3), 904-910.

32. Chetverikov, A., Campana, G., & Kristjánsson, R. (2016). Building ensemble representations: How the shape of preceding distractor distributions affects visual search. *Cognition*, 153, 196–210. https://doi.org/10.1016/j.cognition.2016.04.018

33. Hansmann-Roth, S., Thorsteinsdóttir, S., Geng, J., & Kristjánsson, R. (2021). Temporal integration of feature probability distributions in visual working memory. *Journal of Vision*, 21(9), 1969. https://doi.org/10.1167/jov.21.9.1969

34. van Rooij, I., & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science*, *16*(4), 682-697.

35. 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

36. Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34(4), 273–286. https://doi.org/10.1037/h0070288

37. Swets, J. A. (1986). Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological Bulletin*, 99, 181–198.

38. Luce, R. D., & Galanter, E. (1963). Psychophysical scaling. *Handbook of mathematical psychology*, *1*, 245-307.

39. Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317-1323.

40. Stevens, S. S. (1936). A scale for the measurement of a psychological magnitude: loudness. *Psychological Review*, 43(5), 405–416.

41. Wickens, T. D. (2001). *Elementary signal detection theory*. New York, NY, US: Oxford University Press.

42. Wixted, J. T. (2020). The forgotten history of signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(2), 201–233. https://doi.org/10.1037/xlm0000732

43. Brady, T. F., Schacter, D. L., & Alvarez, G. (2018). The adaptive nature of false memories is revealed by gist-based distortion of true memories. *PsyArxiv*, https://doi.org/10.31234/osf.io/zeg95

44. Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: where next? *Trends in Cognitive Sciences*, 10(7), 292–293. https://doi.org/10.1016/j.tics.2006.05.008

45. Hemmer, P., & Steyvers, M. (2009). A Bayesian Account of Reconstructive Memory. *Topics in Cognitive Science*, 1(1), 189–202. https://doi.org/10.1111/j.1756-8765.2008.01010.x

46. McCarley, J. S., & Benjamin, A. S. (2013). Bayesian and signal detection models. In *The Oxford Handbook of Cognitive Engineering*.

47. Hintzman, D. L. (1986). “Schema abstraction” in a multiple-trace memory model. *Psychological Review*, 93(4), 411–428. https://doi.org/10.1037/0033-295x.93.4.411

48. Howard, M. W., & Kahana, M. J. (2002). A Distributed Representation of Temporal Context. *Journal of Mathematical Psychology,* 46(3), 269–299. https://doi.org/10.1006/jmps.2001.1388

49. Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, 89(6), 609–626. https://doi.org/10.1037/0033-295x.89.6.609

50. Reder, L. M., Nhouyvanisvong, A., Schunn, C. D., Ayers, M. S., Angstadt, P., & Hiraki, K. (2000). A mechanistic account of the mirror effect for word frequency: A computational model of remember–know judgments in a continuous recognition paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(2), 294.

51. Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, 4(2), 145–166. https://doi.org/10.3758/bf03209391

52. Kriegeskorte, N., & Wei, X. X. (2021). Neural tuning and representational geometry. *Nature Reviews Neuroscience*.

53. Xiong, H.D, & Wei, X.X. (2022). Optimal encoding of prior information in noisy working memory systems. Accepted *to Conference on Computational Cognitive Neuroscience (CCN2022).*

54. 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

55. Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. *Advances in neural information processing systems*, 59-68.

56. Shamir, M. (2014). Emerging principles of population coding: in search for the neural code. *Current opinion in neurobiology*, *25*, 140-148.

57. Averbeck, B. B., Latham, P. E., & Pouget, A. (2006). Neural correlations, population coding and computation. *Nature Reviews Neuroscience*, 7(5), 358–366. https://doi.org/10.1038/nrn1888

58. Bartolo, R., Saunders, R. C., Mitz, A. R., & Averbeck, B. B. (2020). Information-Limiting Correlations in Large Neural Populations. *The Journal of Neuroscience*, 40(8), 1668–1678. https://doi.org/10.1523/jneurosci.2072-19.2019

59. Kohn, A., Coen-Cagli, R., Kanitscheider, I., & Pouget, A. (2016). Correlations and Neuronal Population Information. *Annual Review of Neuroscience*, 39(1), 237–256. https://doi.org/10.1146/annurev-neuro-070815-013851

60. Williams, J. R., Robinson, M. M., Schurgin, M.W., Wixted, J.T., and Brady, T.F. (2022). You can't "count" how many items people remember in working memory: The importance of signal detection-based measures for understanding change detection performance. *Journal of Experimental Psychology: Human Perception and Performance*.

61. Robinson, M. M., Benjamin, A. S., & Irwin, D. E. (2020). Is there a K in capacity? Assessing the structure of visual short-term memory. *Cognitive Psychology*, 121, 101305. https://doi.org/10.1016/j.cogpsych.2020.101305

62. 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.

63. 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

64. VanderWeele, T. J., & Mathur, M. B. (2019). Some desirable properties of the Bonferroni correction: is the Bonferroni correction really so bad?. *American journal of epidemiology*, *188*(3), 617-618.

65. Rahnev, D., Block, N., Denison, R. N., & Jehee, J. (2021). Is perception probabilistic? Clarifying the definitions. *PsyArXiv*

66. Eckstein, M. P. (2017). Probabilistic computations for attention, eye movements, and search. *Annual review of vision science*, *3*, 319-342.

67. Ma, W. J. (2012). Organizing probabilistic models of perception. *Trends in cognitive sciences*, *16*(10), 511-518.

68. Zeng, T., Tompary, A., Schapiro, A. C., & Thompson-Schill, S. L. (2021). Tracking the relation between gist and item memory over the course of long-term memory consolidation. *ELife*, 10. https://doi.org/10.7554/elife.65588

69. Rosenbaum, D., & Bowman, H. (2021). Extraction of gist without encoding of individual items in RSVP of numerical sequences. *OSF.*

70. Hommel, B., Hapman, C. S., Cisek, P., Neyedli, H. F., Song, J. H., & Welsh, T. N. (2019). No one knows what attention is. *Attention, Perception, & Psychophysics*, *81*(7), 2288-2303.

71. Greene, N. R., & Naveh-Benjamin, M. (2021). The effects of divided attention at encoding on specific and gist-based associative episodic memory. *Memory & Cognition*, 1-18.

72. Chen, Z., Zhuang, R., Wang, X., Ren, Y., & Abrams, R. A. (2021). Ensemble perception without attention depends upon attentional control settings. *Attention, Perception, & Psychophysics*, *83*, 1240-1250.

73. Zepp, J., Dubé, C., & Melcher, D. (2021). A direct comparison of central tendency recall and temporal integration in the successive field iconic memory task. *Attention, Perception, & Psychophysics*, *83*(3), 1337-1356.

74. Gershman, S. J. (2021). The rational analysis of memory. In *The Oxford handbook of human memory*. Oxford University Press Oxford, UK.

75. Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(6), 1411–1436. https://doi.org/10.1037/0278-7393.24.6.1411

76. Nadarajah, S., Afuecheta, E., & Chan, S. (2019). On the distribution of maximum of multivariate normal random vectors. *Communications in Statistics-Theory and Methods*, *48*(10), 2425-2445.

77. Li, A. Y., Liang, J. C., Lee, A. C., & Barense, M. D. (2020). The validated circular shape space: Quantifying the visual similarity of shape. *Journal of Experimental Psychology: General*, *149*(5), 949.

Acknowledgements

For funding we acknowledge the National Institute of Health, for the National Research Service Award Fellowship (1F32MH127823-01; for M. M. R.) and the National Science Foundation (BCS-1653457 and BCS-2146988 for T. F. B.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

M. M. R. and T. F. B. contributed to conception and design of experiments, developed material and analytic tools and models, and co-wrote the paper. M. M. R. implemented the main experiments and modeling analyses.

Competing interests

The authors have no financial or non-financial competing interests to report.

Figure legends

***Fig. 1 | Laboratory ensemble tasks**. Examples of laboratory ensemble tasks in which participants are typically asked to report on the average along a stimulus dimension, such as color or shape, using a continuous reproduction task.*

***Fig. 2 | TCC framework for memory of individual items and ensembles** A) The TCC framework merges principles of memory uncertainty and the exponential generalization gradient. On the left panel, the two Guassian distributions represent the distribution of familiarity signals for old (purple color) and new (yellow/orange) items, with values denoting greater familiarity. Purple, having been seen, on average has higher familiarity, but on a given trial people judge just one sample from this distribution, such that sometimes, yellow/orange may feel more familiar. In line with signal detection theory, the distance between these distributions (d') quantifies memory fidelity. The right panel shows the psychophysical similarity function shows how average familiarity scales as a function of psychophysical similarity to the remembered item (e.g., the purple at the center of the distribution has familiarity equal to d', and the orange/yellow at the edges has familiarity equal to zero). B) TCC allows us to link models for individual items and ensembles. The left panel shows the TCC model for individual items, which postulates that each item in the memory array elicits a distributed pattern of activation over features values (upper left), which is corrupted by noise (center left). When tested, people report on the feature value that generates the maximum memory signal based on the probed item (lower left). The right panel shows the TCC Perceptual Summation model for ensembles, which postulates that each item in the memory array elicits a distributed pattern of activation over feature values, which are pooled at an early encoding stage of processing (upper right). This distributed pattern of activation is corrupted by noise (middle right) and when queried on the mean color, people report on the feature value that elicits the maximum memory signal in this ensemble representation (lower right).*

***Fig. 3 | TCC ensemble models**. Schematic of all ensemble models that fall within the TCC framework. Within the TCC framework, all ensemble models posit that activations of individual memory representations, quantified with the signal-to-noise ratio (d') underlie ensemble memory processes. Accordingly, each model can be used to predict ensemble data with zero-free parameters by independently estimating a signal-to-noise ratio from a working memory task and substituting it into the ensemble models. Each model provides a way of linking memory for individual items to memory for ensembles, but each embodies different theoretical assumptions regarding how ensembles are computed. The Perceptual Summation model postulates that each item elicits item-specific patterns of activation that are pooled at an early encoding stage of processing. This model has the potential to capture key predictions in the ensemble memory literature, such as that ensemble representations are more robust than representations for individual items, but that item-specific memory can still predict aspects of ensemble memory. The Post-perceptual Summation model, in contrast, captures the view that representations of individual items are maintained in working memory until memory is probed. Finally, the Automatic Averaging model postulates that ensemble representations of the average are extracted automatically, rather than being built up from representations of individual items. Together, comparing these models allows us to formally test key predictions in the ensemble literature using formal model comparison as well as a principled, theoretical framework of memory processes.*

***Fig. 4 | Non-TCC ensemble models**. Schematic of all ensemble models that do not fall within the TCC framework. The non-TCC models differ in the extent to which they represent uncertainty (noisy point estimate; average of point estimates; precision point estimate) versus do not (noise-free point estimate) and the extent to which they assume an averaging of individual (average of point estimates) versus a direct extraction of the ensemble mean (noise-free and noisy point estimate; precision point estimate). See main text for full model descriptions.*

Fig. 5 | Color, shape and sequential memory tasks. *The top two panels show example trial sequences used in the visual working memory (VWM) and ensemble tasks with color (A) and shape (B). In the visual working memory tasks (left), participants saw a set of colors (top) or shapes (middle) and then after a delay a single location was probed, and participants had to indicate which color or shape was in that position. In the ensemble tasks (right), participants saw a set of colors (top) or shapes (middle) and then after a delay were probed on the mean color or mean shape (e.g., a summary of the entire set) rather than on a single individual item. The bottom panel shows an example trial sequence used in the visual working memory and ensemble task in Experiment 5 (C). On each trial of the visual working memory task (left), participants saw a number of colored real world objects presented one at a time and then a single object appeared at test, in grayscale, and participants had to indicate what color that particular item had been. In the ensemble task (right), participants saw a sequence of colors on a single real-world object, and at test had to indicate the average color of this object.*

Fig. 6 | Comparison in predictive accuracy between Perceptual Summation model and competing models of ensemble memory for color with the set size manipulation. *The top panel shows violin plots based on the difference in predicted negative log likelihood scores between each of the six alternative competing models ($PNLL_{Alt}$) and the main Perceptual Summation model ($PNLL_{PerSum}$) for Experiment 1 (n= 50 participants). Lower values of PNLL indicate higher predictive accuracy, therefore, PNLL difference scores higher (or lower) than zero indicate support for the Perceptual Summation (or a competing) model. In both experiments, the vast majority of participants are better predicted by the Perceptual Summation model than any of the alternatives. The bottom panel shows a table with a summary of descriptive and inferential statistics from all comparisons in Experiment 1, including the mean and standard error of the mean across participants. PNLL values were compared with a paired two-tailed t-test, corrected for multiple comparisons and all p-values were statistically significant (p < .001).*

Fig. 7 | The Perceptual Summation model predicts ensemble memory for color with a set size manipulation. *Graphical representation of TCC models' fit and prediction of data in Experiments 1. In this experiment participants had to remember colors of simultaneously presented circles, and the number of colors was manipulated in the working memory and ensemble task. The top row shows the fits of the TCC model for individual items to aggregate data from the visual working memory task for six items. d' estimates from the visual working memory task were substituted into the TCC Perceptual Summation (blue), Post-perceptual (red) and Automatic Averaging (green) models to predict the ensemble data. The bottom panel shows model predictions for a few example participants. We visually show the fits of the TCC model for individual items to the visual working memory data to demonstrate that it provides a reasonable fit to the data (for extended model comparison between this model and other contending models for individual items based on fit and predictive accuracy see: Schurgin et al., 2020).*

Fig. 8| Comparison in predictive accuracy between Perceptual Summation model and competing models of ensemble memory for color with color range manipulation. *The top panel shows violin plots based on the difference in predicted negative log likelihood scores between each of the six alternative competing models ($PNLL_{Alt}$) and the main Perceptual Summation model ($PNLL_{PerSum}$) for Experiment 1 (n=50 participants). Lower values of PNLL indicate higher predictive accuracy, therefore, PNLL difference scores higher (or lower) than zero indicate support for the Perceptual Summation (or a competing) model. In both experiments, the vast majority of participants are better predicted by the Perceptual Summation model than any of the alternatives. The bottom panel shows a table with a summary of descriptive and inferential statistics from all comparisons in Experiment 2, including the mean and standard error of the mean across participants. PNLL values were compared with a paired two-tailed t-test, corrected for multiple comparisons and all p-values were statistically significant (p < .001).*