A Population Response Model of Ensemble Coding

Igor S. Utochkin[1], Jeunghwan Choi[2], & Sang Chul Chong[2,3]

[1]HSE University, Russia

[2]Graduate Program in Cognitive Science, Yonsei University

[3]Department of Psychology, Yonsei University

Corresponding authors:

Igor S. Utochkin
e-mail: isutochkin@gmail.com

Sang Chul Chong
e-mail: scchong@yonsei.ac.kr

Abstract
Ensemble representations have been considered as one of strategies that the visual system adopts to cope with its limited capacity. Thus, they include various statistical summaries such as mean, variance, and distributional properties and are formed over many stages of visual processing. The current study proposes a population coding model of ensemble perception to provide a theoretical and computational framework for these various facets of ensemble perception. The proposed model consists of a simple feature layer and a pooling layer. We assumed ensemble representations as population responses in the pooling layer and decoded various statistical properties from population responses. Our model successfully predicted averaging performance in orientation, size, color, and motion direction across different tasks. Furthermore, it predicted variance discrimination performance and the priming effects of feature distributions. Finally, it explained the well-known variance and set size effects and has a potential for explaining the adaptation and clustering effects.

Keywords: *Ensemble representation; Population code; Variance effect; Set size effect*

Visual environment is composed of complex and dynamic objects, their relationships, and spatial layout. We stroll in a botanical garden and enjoy the spring blossom of flowers without bumping into a group of people passing by. Given the limited capacity of our visual system (Broadbent, 1958; Luck & Vogel, 1997; Palmer et al., 2011), how can we process the color of many flowers and the direction of people at the same time? Ensemble summaries provide one solution for the visual system to deal with this dilemma. The visual system does not have to represent the complex information in detail. Rather, it can summarize similar and redundant information as statistical properties to manage it efficiently within the limited capacity. Statistical summaries include average information, variability, and distributional properties. They are extracted and formed in multiple stages of visual processing and useful for many visual functions.

Average information is extracted from ensembles of similar objects more readily than their constituents, even if the exact average is not physically presented. When multiple items are present and over the capacity limit of the visual system, observers can judge the average information more accurately than features of individual items (Ariely, 2001; Alvarez & Oliva, 2008; Haberman & Whitney, 2007; Parkes et al., 2001; Robitaille & Harris, 2011). Observers are likely to report having seen an item in a set, if it is similar to the average of the set (Ariely, 2001; Khayat & Hochstein, 2018, 2019). This tendency increases as a function of its distance to the mean. Moreover, the memory of individual items is biased toward their average (Brady & Alvarez, 2011; Corbett, 2017; Utochkin & Brady, 2020; Son et al., 2020), although they are not completely replaced with the average. Despite the bias, individual items are remembered more precisely if they are similar (Utochkin & Brady, 2020; Son et al., 2020). Therefore, the average information seems to be the best approximation of a set's gist that is represented with relatively good precision and can compensate for less precisely represented constituents.

To represent a complex scene accurately, however, statistical summaries should include more than just a single representative value like average. For example, we consider a variety of statistical properties, when we are shopping for a pack of cherry tomatoes in a farmer’s market. The average color of cherry tomatoes indicates how ripe they are, and color variability indicates how homogeneous their ripeness is. We also consider their average size, size variability, and size distribution to choose the right pack. Accordingly, the visual system can accurately compute average color (Virtanen et al., 2020), color variability (Maule & Franklin, 2020), average size (Chong & Treisman, 2003), size variability (Solomon et al., 2011), and even richer information about the entire distribution of sizes (Kim & Chong, 2020) and colors (Chetverikov et al., 2017).

Moreover, the single mean for all items is not always a good representative of these items: for example, when yellowish flowers and greenish leaves are presented together on the same tree, their normal ranges of color variation are so different that there is of no use to calculate their common mean but it is better to get local means for each type of items. It appears that whether all items are considered a single group with one mean or as different groups with no single mean depends on the shape of a feature distribution (Treue et al., 2000; Utochkin, 2015). When the features are overall quite similar and smoothly distributed they are likely to form a single-peak distribution with one mean as a good representative of the whole set. But when there are a lot of highly dissimilar features and nothing in-between them, the distribution would have several peaks corresponding to more local groups. This suggests that the nuanced information about the feature distribution can be also an important aspect of ensemble representations.

Ensemble processing is involved with multiple stages of visual processing. Ensemble representations are accurately formed for low-level features such as orientation (Dakin & Watt, 1997), motion direction (Watamaniuk et al., 1989), speed (Watamaniuk & Duchon, 1992), color (Virtanen et al., 2020), brightness (Bauer, 2009), location (Alvarez & Oliva, 2008), and size (Ariely, 2001). They are also formed for high-level features such as facial expression (Haberman & Whitney, 2007), identity (Roberts et al., 2019), lifelikeness (Yamanashi Leib et al., 2016), and economic value (Yamanashi Leib et al., 2020). They are formed more accurately if included items are similar to each other (Ariely, 2001; Corbett et al., 2012; Dakin, 2001; Im & Halberda, 2013; Maule & Franklin, 2015; Sweeny et al., 2013; Solomon, 2010; Solomon et al., 2011; Utochkin & Tiurina, 2014) and if more items are included during averaging due to noise cancellation (Allik et al., 2013; Baek & Chong, 2020a; Brezis et al., 2018; Haberman & Whitney, 2010; Lee et al., 2016; Parkes et al., 2001; Robitaille & Harris, 2011; Solomon, 2010; Solomon et al., 2011)[1].

Ensemble representations are ubiquitous and useful for many important visual functions. They can be used for learning and identifying categories (Khayat & Hochstein, 2019; Utochkin, 2015). The average information of a set can be considered as its prototype and its range can indicate a category boundary (Duffy et al., 2010; Khayat & Hochstein, 2019; Khayat et al., 2021). Previous ensemble representations are linked to current ensembles and individual representations to promote visual stability (Manassi et al., 2017). People use the distribution of distractor properties to find a target (Chetverikov et al., 2016; Treisman & Gormican, 1988;

1 Note that ensemble representations can become less precise with more number of items, when included items are dissimilar (e.g., Ji & Pourtois, 2018).

Utochkin & Yurevich, 2016). Stable ensemble representations facilitate visual search, even when they are not predictive of a target location (Corbett & Melcher, 2014). Ensemble representations are formed based only on a surface information automatically excluding contour information embedded in the surface (Cha & Chong, 2018), suggesting that ensemble representations can be used for figure/ground segregation. Finally, they are useful for texture perception (Cavanagh, 2001) like judging surface qualities such as glossiness (Motoyoshi et al., 2007).

Existing observer models of ensemble perception provide strong quantitative predictions that fit the performance of human observers well in some ensemble tasks. Most of the models account for averaging (Allik et al., 2013; Baek & Chong, 2020a; Parkes et al. 2001; Solomon et al., 2011) and only few of them go further to account for variability perception (e.g., Morgan et al., 2008; Solomon, 2010). These models are focused on quantifying parameters of internal representations that underlie the observed performance. For example, the precision of average computation was shown to depend on early noise involved with individual representations, the number of averaged items (Allik et al., 2013; Baek & Chong, 2020a; Parkes et al. 2001; Solomon et al., 2011), late noise involved with average computation (Baek & Chong, 2020a; Parkes et al. 2001; Solomon et al., 2011), and distributed attention (Baek & Chong, 2020a). Whereas these models predict average and variance discrimination performances well (Allik et al., 2013; Baek & Chong, 2020a; Parkes et al. 2001; Solomon, 2010; Solomon et al., 2011), they did not explain the representation of other distributional properties. More importantly, they do not explain the core computational mechanism behind the perception of average and variance. They implicitly assume that the computational process behind average and variance perception is akin to the calculating arithmetic mean and variance of features respectively. For example, the arithmetic mean is defined as the sum of all sampled measurements divided by their number. However, it is not likely that the visual system does exactly that computation given that the judgments of sum of visible features are noisier than those of the average (Lee et al., 2016; Raidvee et al., 2021). Moreover, these models do not specify the underlying neural mechanism of arithmetic mean and variance computation.

In the current paper, we propose a general, neurally plausible, and mechanistic model of ensemble representations across different feature domains. The model is based on population coding and spatial pooling, two very basic principles of organizing sensory systems of the brain. Many visual features are encoded as population responses, which implies that any particular feature value causes a specific pattern of responses (firing rates) in a group of neurons in accordance with their tuning to that feature (Brouwer & Heeger, 2009; Ester et al., 2009; Georgopoulos et al., 1986; Treue et al., 2000). Spatial pooling occurs from the hierarchical

organization of visual processing that involves the progressively increasing size of receptive fields (areas of the visual field to which the neuron is responsive) from lower to higher brain regions (Boussaoud et al., 1991; Desimone & Schein, 1987; Ungerleider & Bell, 2011). Although previous studies have considered a population response model to explain ensemble perception (Baek & Chong, 2020b; Chong & Treisman, 2003; Haberman & Whitney, 2012; Hochstein et al., 2018), these were descriptive outlines that have not been implemented as computational models, and thus they have not been tested against actual data. One exception is a population coding model of number averaging (Brezis et al., 2018), but even this model was confined to explain temporal averaging of numbers without generalization to other statistical properties or features. Furthermore, this model's averaging process is the mathematical abstraction of population responses to numbers without providing the plausible neural mechanism for it. Their model employs multiple population responses of number-selective neurons, each population responding to an individual presented number, and a pooling layer to decode an average value from these responses using a parabolic tuning curve. We propose an alternative version of the population coding model that focuses on a key role of the pooling layer in ensemble perception. In our model, the pooled population response collected from multiple local responses to individual features is *the full-fledged neural representation of the ensemble* that preserves its rich statistical properties. This pooled response is sufficient to explain human performance in averaging, variance discrimination, and distribution-shape effects. In addition, our proposed model simulated well-known findings in the field of ensemble perception such as over-representation of mean information, the variance effect on averaging, outlier detection, and noise cancellation in large sets.

Methods

The architecture of the Core model

Our Core model of ensemble representation is an algorithm that simulates a pooled population response to a stimulus set of individual features; we operationalize this population response as a neural code of the ensemble. The core model can be presented as a two-layer neural network, with the bottom layer representing individual items in small spatially separated receptive fields and the top layer pooling signals from the bottom layer within a single receptive field (Figure 1). Units of these two layers are feature-selective neurons tuned to fire at a certain rate depending on stimuli. Connections between the layers are feedforward and each bottom-layer unit is connected to each top-layer unit with a certain synaptic weight. The distributions of synaptic weights can be conceptualized as tuning functions of the pooling top-layer units. In a

fully biologically plausible system, stimuli would cause population response in both layers, so that a number of neurons is differentially activated when a given feature is presented. However, for simplicity, we conceptualize responses of the bottom layer as singular feature values corresponding to the actual stimulus value (corrupted by some amount of early noise). We can make this simplification assuming that population responses of a lower layer are already included in measurable outputs of the top-level units, which are used to plot tuning curves. For example, the tuning curve of monkey V4 can be built based on direct measurements from V4 neurons, however, the firing rate of each individual V4 neuron in fact results from the population activity of multiple lower-level neurons. Researchers usually omit the hidden population activity of lower layers, so the tuning curve of the measured neuron usually is plotted directly as a function of stimulus. Therefore, we omit the population responses of the bottom-layer units because we use the same logic.

The wiring scheme of connections between the bottom and the top layers is critical to understand the machinery of the pooling mechanism that creates the neural representation of the ensemble. This wiring scheme defines feature preferences (tuning curves) of the top-layer neurons. In Figure 1, it is shown as a set of synaptic weights (we illustrate them with different numbers). In our example, the maximum weight is 3 that shows perfect matching between the activated bottom-level and top-level units; weights 2, 1, and 0 show response attenuation as a function of decreasing similarity between the preferred and actually presented values (Note that the synaptic weights as well as the shape of the tuning function in Figure 1 are arbitrary and used only for illustration purposes). As the top-layer units are pooling, we can present the outcome activation of each such neuron as the average of synaptic weights produced by each active bottom-layer unit. Note that signal averaging is a neurally plausible way of pooling local responses to different competing visual features by neurons with large receptive fields when no stimulus is selectively attended (Desimone & Duncan, 1995; Kastner et al., 1998). The distribution of individual top-layer outcomes defines the shape of the pooled population code corresponding to an ensemble representation.

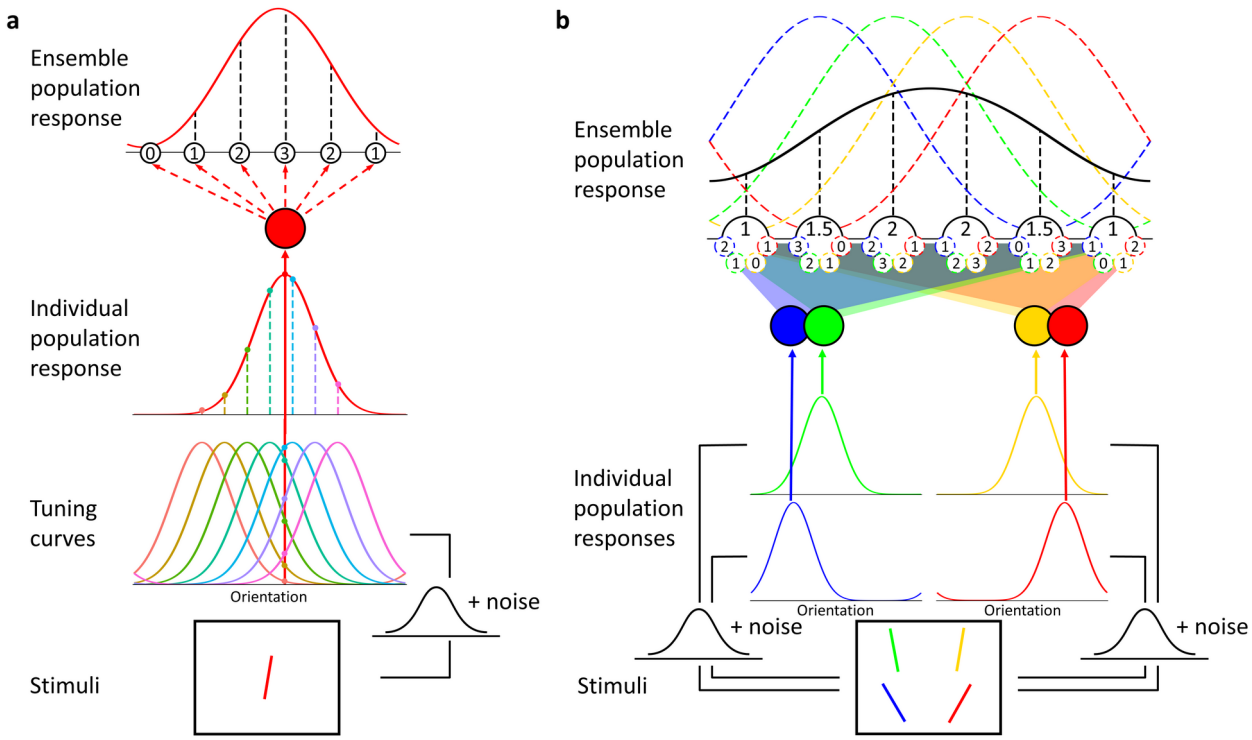


Figure 1. A model of pooling and population coding explaining the formation of a neural representation of ensemble. Panel (a) depicts the case of a single oriented line and panel (b) depicts a set of oriented lines. The bottom layer of the neural network depicts simplified local population responses (each unit is a noisy population code of a single item). The top layer is a population of selectively tuned neurons pooling local signals from the bottom-layer units. The distribution of synaptic weights corresponds to tuning curves of the top-layer neurons. Outputs of the top-layer neurons are the sums of contributing bottom-layer inputs with their synaptic weights. The shape of the population response (neural representation of an ensemble) following from the distribution of top-layer outputs is depicted on the very top of each panel.

Computational modeling

Computational modeling of ensemble perception was implemented in codes written in R language. We have run a set of simulations of various studies from the ensemble literature using different tasks, feature dimensions, and feature distributions. Each simulation was based on the Core model producing a pooled population response and also included a specific part depending on a simulated paradigm. The specific part in most cases included reading-out task-relevant output parameters of the population code, response generation based on these parameters, and performance estimation (e.g., various accuracy metrics – percent correct, just noticeable difference, and *d*'). Model performance was then tested for correlation against the data from the corresponding studies in order to estimate how well the model predicts the observed patterns. The only exception was the feature distribution learning (FDL) paradigm (Chetverikov et al.,

2016, 2017) where population codes produced by the Core model were directly compared against the behavioral data, because the paradigm does not imply any explicit report about ensemble per se (see detailed explanation in Modeling feature distribution learning section of Methods).

For user convenience, most of our simulations were implemented as a set of functions using only stimulus settings as arguments. These functions are collected in the R package, *'EnsPopulationModel'* that can be found at https://osf.io/9yxv5/.

Computational implementation of the Core model

The Core model was computationally implemented as a function (*'popcode'* function) using a vector of feature values of a stimulus set and a feature space specification as arguments. The output of the function is the distribution of spike rates among a set of feature selective neurons x ordered by their feature preference whose preference range and granularity are predefined in a separate numeric vector.

The bottom layer of the Core model implemented in *'popcode'* function takes each individual value from the input stimulus set and adds a random number from a truncated by 2 standard deviations normal distribution with the mean = 0 and standard deviation = σ_n, where σ_n is the amount of early encoding noise. The resulting number is termed *L1*. This models the occurrence of noisy responses to individual items in neural populations with small receptive fields (roughly, V1-like neurons). Each *L1* is passed to the top layer.

The top layer of the model pools the noisy signals of the bottom level. For each individual bottom-layer signal, a population response is produced in all top-layer neurons in accordance with synaptic weights approximated by a normal (for linear dimensions) or wrapped normal (for circular or semi-circular dimensions such as orientation, color hue, and motion direction) probability density functions (f).

For linear dimensions:

$$R(x) = f_{\text{Normal}}(x, L1, SD_{\text{tuning}}) , \qquad (1a)$$

where $R(x)$ are the responses of neurons from a population of neurons x ordered by their stimulus preference to a bottom-level signal $L1$, and SD_{tuning} is a standard deviation of the Gaussian tuning curve of the top-layer neurons.

For circular and semicircular dimensions:

$$R(x) = f_{\text{WrappedNormal}}(x, L1, SD_{\text{tuning}}) \quad (1b)$$

Note that for the circular and semicircular dimensions, x, $L1$, and SD_{tuning} are angular variables entered in degree units. To map these values onto the circular dimension, they were converted to radians as $\theta° \times \pi/180$, where $\theta°$ is the degree value of an argument variable. To map these values onto the semicircular dimension, we used the conversion formula of $\theta° \times \pi/90$.

Since population responses are usually measured in the units of spikes per second rather than probability density, we scaled the R(x) values, so that:

$$R`(x) = R(x) \times maxspike / max(R(x)) \ , \quad (2)$$

where *maxspike* is the spike rate of a neuron at the peak of its feature preference. Since *maxspike* is a constant scaling factor, it can be set arbitrarily to fit neurally plausible values (we set it at 30 Hz in our simulations) but it does not affect the shape of the population response.

The $R`(x)$ along the whole population $x$ can be interpreted as a population response to an individual stimulus if this stimulus is presented alone. In terms of the joint (ensemble) population response, $R`(x)$ is its individual component. Therefore, the contributions of all individual components to the activation of the top-level response will be defined by a matrix $C = \| R`_i(x_j) \|$, where $R`_i(x_j)$ is the i-th component activation of the j-th top-layer neuron.

The pooled response is accomplished by averaging all n components for each of the m top-layer neurons:

$$L2(x) = \left[\frac{\sum_{i=1}^{n} R'(x_1)}{n}, \frac{\sum_{i=1}^{n} R'(x_2)}{n}, \ldots, \frac{\sum_{i=1}^{n} R'(x_m)}{n} \right] \quad (3)$$

Figure 2a shows four illustrative examples of the pooled population code resulting from our Core model as a function of the feature distribution in a stimulus set with different orientations. These four examples demonstrate how the population response takes a shape isomorphous to the feature distribution and thus conveys its critical properties. Green, red and blue curves in Figure 2a show population responses to smoothly distributed orientations with different means and variability. Most notably, middle neurons accumulate more local signals, as they are activated by features from both sides of the synaptic circuit, whereas neurons at edges are activated only from one side. The overall distribution does, therefore, have a single peak around the mean prevailing over the activity of other neurons. The comparison between red and blue curves in Figure 2a illustrates how an ensemble representation depends on feature

variability in a display. When a greater range of features is involved, activation spreads to a larger population of pooling units, so the bandwidth of an active population gets broader as well. Finally, a black curve in Figure 2a shows a population response to a display with highly dissimilar features. Here, clusters of similar features (e.g., all steep lines) contribute to the activation of local subpopulation of pooling neurons but the absence of features in the middle of the physical feature distribution prevents the formation of a single peak in the middle of the population code. Instead, two local peaks are formed that can correlate with the perception of two highly distinct groups of features rather than a single set (Treue et al., 2000; Utochkin, 2015).

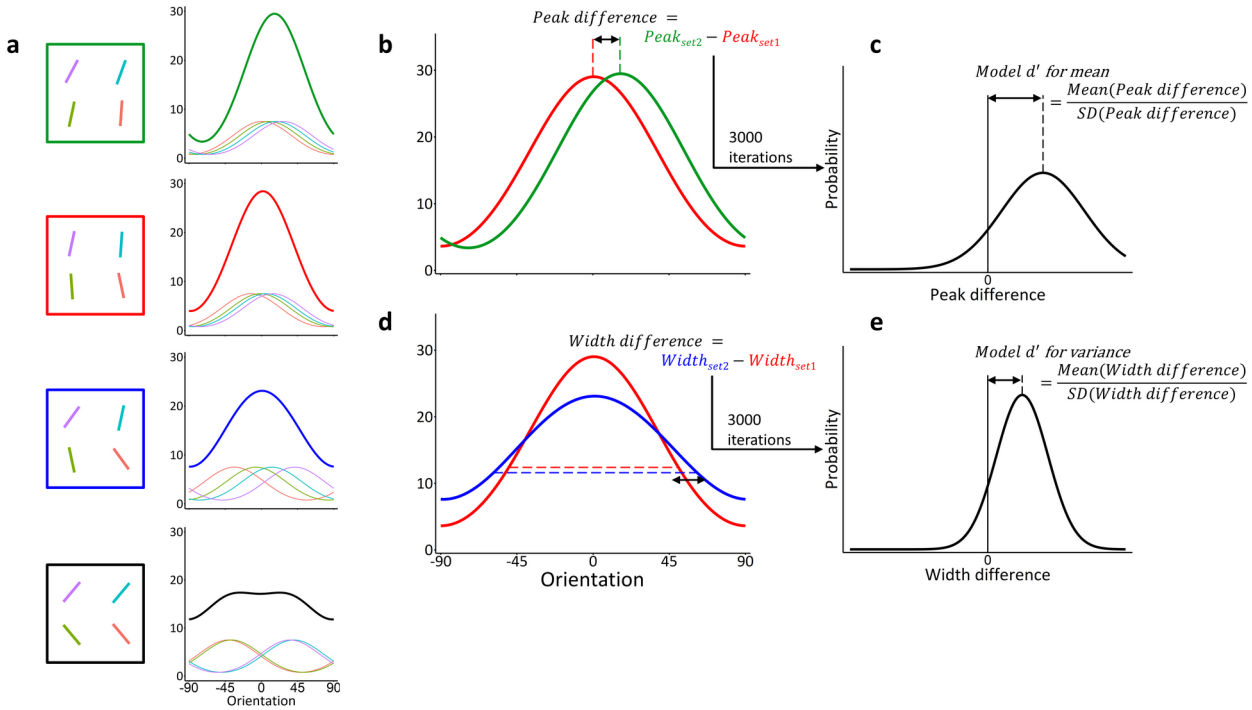


Figure 2. Population responses of the pooling layer (V4-like neurons) to 4-item ensembles of oriented lines: (a) narrow uniform distribution (green & red), broad uniform distribution (blue), and two-peak distribution with extremely spaced orientation values (black). The pooled population response is shown by the thick lines and corresponds to the ensemble neural representation; the thin colored lines depict component population responses to individual stimuli (amplitudes divided by set size due to averaging). The peaks of individual curves are jittered because of the early noise. Our model computes a peak difference from two population responses (b) to simulate a *Model d'* for mean discrimination (c). Similarly, it computes a width difference from two population responses (d) to simulate a *Model d'* for variance discrimination (e).

Output model parameters

The resulting population code *L2* from equation 3 has a set of output parameters that can have straightforward interpretation in terms of ensemble summary statistics that observers might

estimate in a stimulus. The *peak location, P* can be naturally interpreted as the most likely estimate of an ensemble mean. It is defined as the feature preference of neuron having the maximum firing rate *L2* in the population *x*:

$$P = \text{argmax}\ L2(x) \tag{4}$$

The bandwidth of the population activity, calculated as the standard deviation of the involved pooling neurons weighted by their outcome firing rates, can be used as an estimate of ensemble variability. We term this parameter a *population width (W)*. For linear feature dimensions, the *W* was calculated as a weighted standard deviation and specified as W_{linear} (that in our code was implemented by the function 'w.sd' from the R package 'weighted.desc.stat': https://cran.r-project.org/web/packages/Weighted.Desc.Stat/index.html):

$$W_{linear} = \sqrt{\sum_{i=1}^{n} \frac{L2_i \times (x_i - \bar{x})^2}{\frac{n'-1}{n'} \sum_{i=1}^{n} L2_i}} \tag{5}$$

where x_i is the feature preference of the *i*-th neuron, $L2_i$ is the output firing rate of the *i*-th neuron, $\bar{x}$ is the weighted average preferred feature of the population, *n* is the number of neurons in the population, *n'* is the number of non-zero output firing rates.

For circular and semi-circular feature dimensions, the *W* was defined as a weighted circular deviation. The weighted circular deviation is based on the length of the mean resultant vector *r*:

$$r = \frac{\sqrt{\left(\sum_{i=1}^{n} L2_i \cdot \cos x_i\right)^2 + \left(\sum_{i=1}^{n} L2_i \cdot \sin x_i\right)^2}}{\sum_{i=1}^{n} L2_i} \tag{6}$$

Then, the $W_{circular}$ is calculated using the equation for the circular deviation (Berens, 2009):

$$W_{circular} = \sqrt{-2\ln(r)} \tag{7}$$

In our computational model package, two functions are used to calculate the peak location *P* and the population width *W*. The first function, *'decode'* has feature values of a stimulus set as an argument and returns the *P*, the *W*, as well as the real mean feature and real standard deviation of the set. The second function, *'decode.direct'* has a pooled population response (the output of the Core model, *L2*) as an input and returns the *P* and the *W*.

Modeling mean and variance discrimination in 2-AFC tasks

In each trial of a typical two-alternative forced choice task (2-AFC) or two-interval forced choice task (2-IFC), observers are shown two sets of items either in two distinct locations simultaneously or in the same location serially. The observers are instructed to determine the direction of difference between the sets along a target dimension. For example, in the orientation averaging task, observers are asked to answer whether set 2 is tilted more clockwise or counterclockwise than set 1 on average (e.g., Solomon, 2010) or vertical (e.g., Dakin, 2001; Yashiro et al., 2020). In a variance discrimination task, observers are instructed to answer whether set 2 has greater or smaller feature variation than set 1. Regardless of the particular task, the physical difference between the sets is systematically manipulated across trials that allows to estimate the discrimination threshold (also referred to as the just noticeable difference, JND) via the proportion of "set 2 greater" against "set 1 greater" answers as a function of the physical difference.

Simulating observer behavior from pooled population responses

In each run of our simulations, we generated two vectors of values corresponding to individual items in sets 1 and 2. Set sizes and distributional properties of sets were generated in accordance with methods described in the relevant published studies. The absolute physical differences between sets 1 and 2 ranged from 0 to a magnitude at least twice as large as the highest JND value reported in a referred study or set of studies. These two sets of stimulus values were the model input. A random number drawn from a Gaussian distribution (mean = 0, different *SD*s depending on feature dimension) was then added to each individual value to simulate stimulus corruption by the early noise in layer 1. Pooled population responses, as described in our Core model, were then created for each of the sets. The tuning curves of pooling neurons were Gaussian functions with a *SD* depending on feature dimension.

When the population responses for sets 1 and 2 were created their output parameters could be read out to produce an answer on a given trial. For the averaging tasks, a difference between peak locations of population response 2 and population response 1 was read-out (Figure 2b), whereas for the variance discrimination tasks, a difference between population widths between these two population response was read out (Figure 2d). If the difference between the peaks or between the widths was positive then the model observer answered "set 2 greater"; if the difference was negative the model observer answered "set 2 smaller"; if the difference was zero the model observer chose one of these responses randomly. After running 3,000 simulations on each condition (each combination of feature distribution, set size, and physical difference between sets 1 and 2), the probability of "set 2 greater" was calculated (Figures 2c and 2e). The

distribution of these probabilities as a function of the physical difference (the so called psychometric function) was fit as a cumulative Gaussian distribution and the JND was determined using the same threshold as in the reference study.

Signal-detection parameters from the model and behavior

The peak location and the bandwidth parameters of the population code normalized by corresponding variability measures make psychophysically meaningful proxies of sensitivity to ensemble summary statistics. For any given stimulus difference either along the average feature or along feature variance, percent correct answers can be converted to a behavioral discriminability index, *d*`, which reflects a signal-to-noise ratio, the mean distance between the representations of the two sets in the feature space normalized by some internal noise (Macmillan & Creelman, 2004). This internal noise corrupts representations of the sets (or of the set and the comparison item) and makes them confused on a certain proportion of trials. Our model can produce an analog of such *d*` for a 2-AFC task. For each of two sets (or for a set and an item), two population codes are generated and the difference between their relevant output parameters – difference between peak locations for the average comparison task (Figure 2b), difference between bandwidths for the variance comparison task (Figure 2d) – is calculated. Depending on whether the resultant difference is positive or negative, a decision is made as to which one of the sets has a greater mean or variance. Since the early noise makes peak and bandwidth parameters fluctuate from trial to trial, the differences between the representations will also fluctuate which will provide a bell-shaped distribution of representational differences across trials (Figure 2c). The mean of this difference distribution can be conceptualized as the neural difference signal between the population responses. The greater the mean is shifted away from 0, the stronger the signal. The *SD* of the difference distribution can be conceptualized as the amount of discrimination noise. The mean difference normalized by the SD (M_{diff}/SD_{diff}) is the signal-to-noise ratio, or the model's *population d*` (Figure 2c). Note that the model's population *d*` is not supposed to equal the behavioral *d*` that can include other sources of uncertainty (e.g., related to decision) not considered by our model. Yet, we consider the introduction of the population *d*` useful as a sensible, measurable concept that provides simple mapping of the population code model onto psychophysical measurements in behavioral experiments.

The behavioral *d*` for a given condition can be obtained from the observed JND assuming that the psychometric function, from which this JND is calculated, is a cumulative normal distribution. Since the JND is the stimulus difference that provides a known probability of correct discrimination, this probability can be *z*-transfromed and the SD of the psychometric

function can be found. Based on the SD, a *z*-score for any physical difference between two sets can be calculated and used to find the 2-AFC *d*`(see the formula in Stanislaw & Todorow, 1999). The distributions of behavioral and model *d*`s across various stimulus conditions and in a fixed physical difference can be further used to estimate their correlation that would show the quality of model predictions regarding behavior.

Modeling averaging in the method of adjustment

In the method of adjustment, observers are typically shown a single sample stimulus (e.g., a set of items) and are asked to set a subsequently presented adjustable probe as close as possible to the target feature of the sample (e.g., mean orientation). Since the probe can be adjusted in an almost continuous fashion, the data are usually presented as a continuous error distribution across trials. The mean of this distribution is taken for the point of subjective equality and its standard deviation is taken as discrimination precision measure (e.g., a JND). Since there is only one set presented, we simulate our model responses in the adjustment method by simply picking a corresponding output parameter (e.g., peak location for averaging) of a simulated pooled response. The average output parameter across runs is taken for the point of subjective equality and the standard deviation is taken for precision (not including late noise factors).

Modeling feature distribution learning

Behavioral data. In feature distribution learning (FDL) experiments, observers are first exposed to a streak of trials with a consistent distractor distribution and an odd-one-out target (for example, a line with a singleton, highly distinctive orientation or a diamond with a highly distinctive hue – Chetverikov et al., 2016, 2017). After priming an observer with such a streak of trials, a switch trial follows where the distractor distribution dramatically shifts its mean feature (and can also change its shape) and the target feature also takes a new value that can fall within the range of the former distractor distribution (hereinafter, the *priming distribution*), as well as beyond this range. Observers' task was always to find the target as fast as possible and to respond as to which part of the array it is located. Following the principal logic of the FDL paradigm, our analysis was focused on reaction times (RT) in the switch trials where priming effects of the previously learned distractor distribution are tested. The distribution of RT's in the switch trials as a function of the target feature distance from the middle-range feature of the former distractor distribution (hereinafter, TD distance) was our main subject of prediction. We

also applied local polynomial regression fits of RT x TD distance functions (same as in Chetverikov et al., 2016, 2017). Correlation coefficients between the RT x TD distance data (binned by 5 degrees) and the regression fits were considered to reflect the "ceiling" quality of the fits that might be reached with a smooth continuous function based only on the data. The correlations between the data and model predictions are further compared with these ceilings.

Physical priming distributions. In order to see how much the RT distributions are predicted by the distributions of physically presented features, we directly used probability density functions as in Chetverikov et al. (2016)'s experiments that we chose for modeling: Gaussian distributions with SD = 10 degrees and SD = 15 degrees, a uniform distribution covering from -30 to 30 degrees (all three distributions tested in Chetverikov et al. (2016)'s Experiment 2), and two skewed, triangular distributions ranging from -30 to 30 degrees, either right-skewed (peak = -25 degrees), or left-skewed (peak = 25 degrees) (these two shapes were tested in Chetverikov et al. (2016)'s Experiment 4). The probability densities based on these theoretical distributions were correlated with the RT x TD distance data in the corresponding priming distribution conditions to estimate how well the RT distributions are predicted by the real feature distributions.

Population model distributions. To estimate how much the RT distributions primed by a certain distractor distribution can be predicted by the ensemble population response, we estimated the correlation between the RT distributions and the average spike rate distributions provided by our core model for corresponding priming distributions. We had 3,000 runs for each type of the priming distribution (as listed in the previous paragraph). In each run, we randomly drew 35 feature values from a defined distribution, each feature value corresponded to one individual distractor from Chetverikov et al.'s (2016) standard displays (they always presented an array of 36 elements, one of them reserved for the target and the rest 35 for distractors). In accordance with our core model, these 35 individual features were corrupted by the early noise and then pooled by neurons with the default tuning parameters. The output spike rate distributions were then averaged across runs to get the model prediction of the ensemble population response.

Note that for visualization purposes we normalized RT's, probability densities of features in physical priming distributions, and spike rates in model's population responses to fit these functions in the same scale from 0 (minimum value) to 1 (maximum value). Note that this normalization does not affect any correlations between the distributions, as it only changes the scale but not relations between the points.

Modeling results

We applied our 'pooling and population coding' model to different paradigms and feature dimensions used in previously published studies to see whether our model can account for the variety of data accumulated in the field. To remind, the model focuses on the relatively early representational mechanism of ensemble perception that is limited by the early input noise and by tuning characteristics of pooling neurons. The model does not include late sources of uncertainty and noise (such as readout error, decision error, or lapses). Therefore, our model naturally does not predict behavioral data exactly. We demonstrate instead that, although model predictions and data may be differently scaled, the model nicely captures basic patterns and trends observed in the data (including some interesting non-linearities and non-monotonicity, such as the "dipper function" in variance perception – see details below). The principal significance of these demonstrations is that a broad range of effects can be captured by exactly the same simple mechanism without introducing any additional assumptions.

Average discrimination in a 2-AFC task. 2-AFC discrimination of mean orientations has been employed in a number of studies (Cha & Chong, 2018; Dakin, 2001; Dakin & Watt, 1997; Solomon, 2010; Rosenholtz, 2000). In some of these works, researchers have parametrically manipulated set size (the number of elements in a set) and the physical variance of sets (e.g., Dakin, 2001; Solomon, 2010). They found the variance and set size effects. The variance effect shows that the mean discrimination threshold, as calculated from the slope of a psychometric function (the proportion of answering that one stimulus is greater than another as a function of the physical difference between them), is plateaued at the beginning (when stimulus variances are up to 3-4 degrees) and then steadily grows as a function of variance (Figure 3a). The variance effect is robust not only in an orientation domain but also in other domains such as color (Maule & Franklin, 2015) and size (Im & Halberda, 2013; Solomon et al., 2011). The set size effect indicates that mean discrimination threshold generally improves with increasing set size (Figures 3a). Such a benefit from larger set sizes can be also found in some of the size averaging experiments (e.g., Allik et al., 2013; Baek & Chong, 2020a; Chong et al., 2008; Robitaille & Harris, 2011), although other studies failed to report it (Ariely, 2001; Marchant et al., 2013). Existing accounts for these findings are somewhat complicated and usually introduce assumptions leading to additional free parameters changing between conditions, such as a variable number of items considered from the entire set (*effective set size*, Allik et al., 2013;

Dakin, 2001; Dakin & Watt, 1997; Im & Halberda, 2013; Maule & Franklin, 2016; Solomon, 2010: Whitney & Yamanashi Leib, 2018) or the variable amount of early noise (Alvarez, 2011).

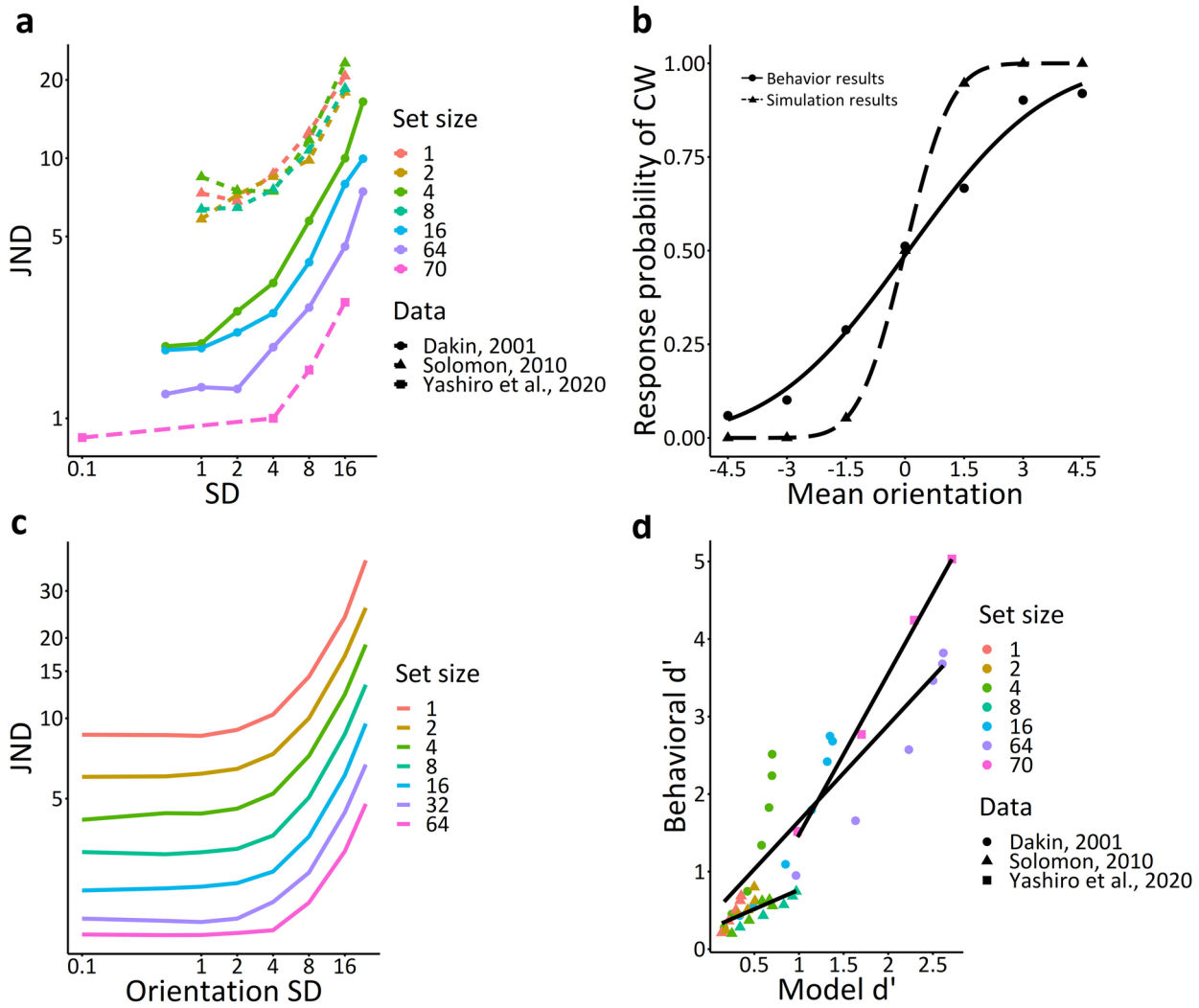


Figure 3. The variance and set size effects on average discrimination in a 2AFC paradigm. (a) Previous studies have found that mean orientation discrimination performance decreased with increasing variance and decreasing set size (Dakin 2001; Solomon, 2010; Yashiro et al., 2020). (b) Psychometric functions of orientation discrimination. The solid line indicates the results from Cha & Chong (2018) and the dotted line is the results of the model simulation. (c) Model JNDs are plotted against SD and set size. They capture the qualitative patterns of behavioral results presented in (a) well. (d) The correlation results between behavioral *d*' from previous studies (Dakin 2001; Solomon, 2010; Yashiro et al., 2020) and model *d*'.

Orientation. Our model explains both the plateau-and-growth variance effect and the facilitating set size effect[2] on the discrimination of average orientations. Noteworthy, the model does it parsimoniously, without the need to fit free parameters to every situation. We used a fixed set of parameters based on the previous research of orientation averaging and orientation population coding. Namely, the width (SD) of the Gaussian tuning curve was set at 37 degrees

2 Our model can explain no effect of set size on orientation averaging, if early noise increases with set sizes (Supplementary Figure 3c).

(McAdams & Maunsell, 1999) and the SD of the Gaussian early noise was set at 4 degrees (Dakin, 2001; Solomon, 2010). Figure 4 shows how our model simulates mean orientation discrimination. When two sets of orientation stimuli are presented (Figure 4a), population responses for each set can be simulated and the peak of each population response can be compared to each other to simulate a trial of mean orientation discrimination (Figure 4b). This procedure was repeated for 3,000 times to make the distribution of population response differences (Figure 4c). Finally, Figure 4d shows how we constructed a psychometric function based on multiple population response differences depending on the mean orientation differences of two sets. Figure 3b shows the simulated psychometric function of Cha & Chong (2018). With the same set of parameters, Figure 3c depicts the simulated JND as a function of orientation variances of both stimuli and set size. Psychometric functions start with a plateau that turns into a steady growth at about ~4 degrees. This is what typically described for the variance effect on mean discrimination (Dakin, 2001; Solomon, 2010). We also observed that the greater the set size, the lower the curve which was consistent with previous studies (Dakin, 2001; Solomon, 2010; Yashiro et al., 2020). Figure 2 shows how these patterns arise from model outputs. For a fixed mean difference between two stimuli, the mean difference between the corresponding population peaks is also fixed but the SD of peak differences grows as a function of the stimulus variance. In contrast, increasing set size reduces the SD of peak differences. The mechanism of this reduction is presumably based on the *wisdom of crowds* (Alvarez, 2011; Galton, 1907), or *noise cancellation* (Baek & Chong, 2020b; Sun & Chong, 2020) that implies that averaging multiple imprecise individual estimates eventually gives more precision because uncorrelated individual errors are cancelled out. Accordingly, the signal-to-noise ratio, or *population d`* decreases with increasing stimulus variance and increases with increasing set size. Thus, our model captures the pattern of data found in previous studies (Dakin 2001; Solomon, 2010; Yashiro et al., 2020). Figure 3d plots data from these studies converted to behavioral *d`* (see Figure 2 for the conversion algorithm) against the population *d`*. Pearson's correlation between the population and behavioral *d`*s was r = .87, .67, and .99 respectively.

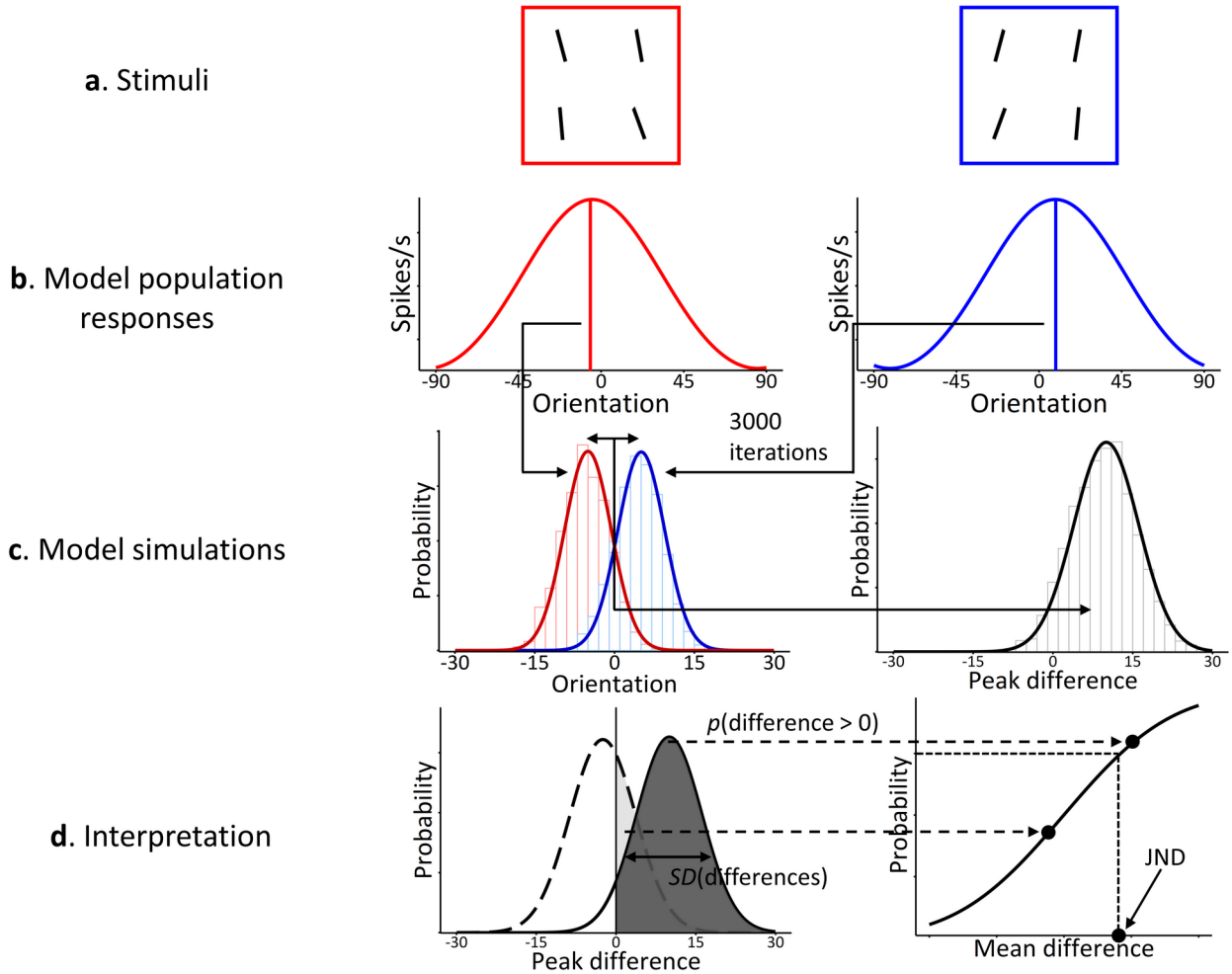


Figure 4. The procedure of mean orientation discrimination by the model. (a) Two sets of oriented bars. The left set had the mean orientation of -5 deg and the right set had the mean orientation of 5 deg. (b) Model population responses to each set of stimuli. (c) We simulated population response differences for 3000 times to make their distribution. (d) Based on these distributions, we constructed a psychometric function and reads a JND from it.

Size. We also applied our model to the results of mean size discrimination (Baek & Chong, 2020a; Solomon et al., 2011). We changed only the standard deviation of tuning curve of pooling neurons to 0.3 degrees in the space where the unit was the diameter with an exponent of 1.52. We used this unit to reflect apparent sizes (Chong & Treisman, 2003; Teghtsoonian, 1965) and chose 0.3 degrees based on psychophysical findings of early noise (Baek & Chong, 2020a). We have successfully described the variance and set size effects in size averaging (Baek & Chong, 2020a; Solomon et al., 2011, Supplementary Figure 1). Peason's correlation between the model *d*`s and behavioral *d*`s was high (r = .98 for Baek & Chong, 2020a, and .88 for Solomon et al., 2011).

Motion direction and color. Here, we applied our model to both motion direction (Watamaniuk et al., 1989) and color averaging (Virtanen et al., 2020). In the case of direction averaging, we set the standard deviation of tuning curve of pooling neurons to 40 degrees with

the early noise of 7 degrees (Albright, 1984; Maunsell & van Essen, 1983; Rodman & Albright, 1987). The standard deviation of tuning curve of pooling neurons was 40 degrees with the early noise of 15 degrees in the case of color averaging (Maule & Franklin, 2016). Supplementary Figures 3 shows the behavioral and simulation results. The model has successfully described the results of direction averaging (Watamaniuk et al., 1989, Supplementary Figure 3a) and color averaging (Virtanen et al., 2020, Supplementary Figure 3b). Peason's correlation between the model and behavioral *JND*'s was high (r = .98 for Watamaniuk et al., 1989, and .81 for Virtanen et al., 2020).

Orientation and size averaging using the adjustment method. The model can also predict the results of averaging studies using an adjustment method (Epstein et al., 2020; Khvostov & Utochkin, 2019; Kim & Chong, 2020; Utochkin & Brady, 2020). We did not change any parameters of the model, and used the peak of population response to a set of presented stimuli to predict an observer's reported mean in each trial. We have successfully described the results of an adjustment method in orientation averaging (Utochkin & Brady, 2020, Epstein et al., 2020, Supplementary Figures 2a and 2b respectively) and in size averaging (Khvostov & Utochkin, 2019; Kim & Chong, 2020, Supplementary Figures 2c and 2d respectively). Pearson's correlation between the predicted and reported means was high (r = .96 for Utochkin & Brady, 2020, .88 for Epstein et al., 2020, r = .55 for Khvostov & Utochkin, 2019, .78 for Kim & Chong, 2020). We also found that our model can predict the variance effect in an adjustment method. As the range of stimuli increased (from red to blue clouds in Supplementary Figures 2a and 2c), clouds of dots dispersed more, showing the variance effect. The difference between the predicted and reported means became larger with larger variance.

Orientation variance discrimination. Although variance discrimination is a less studied topic than averaging, several studies addressed it in the orientation domain (Jeong & Chong, 2020, 2021; Morgan et al., 2008; Norman et al., 2015). Jeong and Chong (2021) found that observers judged a display with orientations concentrated on edges (outward condition) more variable than a display with orientations concentrated on a center (inward condition, Figure 5a) . Since the ranges of two displays were the same, these results suggest that people's variance discrimination is better described by standard deviation than range. Morgan et al. (2008) and Solomon (2010) have systematically measured JND's of variance discrimination as a function of pedestal variance (the lowest variance of the two compared sets). An interesting discovery by Morgan et al. (2008) was the non-monotonic, "dipper" function of variance discrimination. That is, JND's decreased with pedestal variance when it was small (up to ~4 degrees) but the JND's increased when the pedestal variance became larger (Figure 5b). One suggested explanation for

this dip in the JND function is that the visual system distinguishes slight increase of external noise more easily because of a threshold discounting a small amount of internal noise (Morgan et al., 2008; Solomon, 2009).

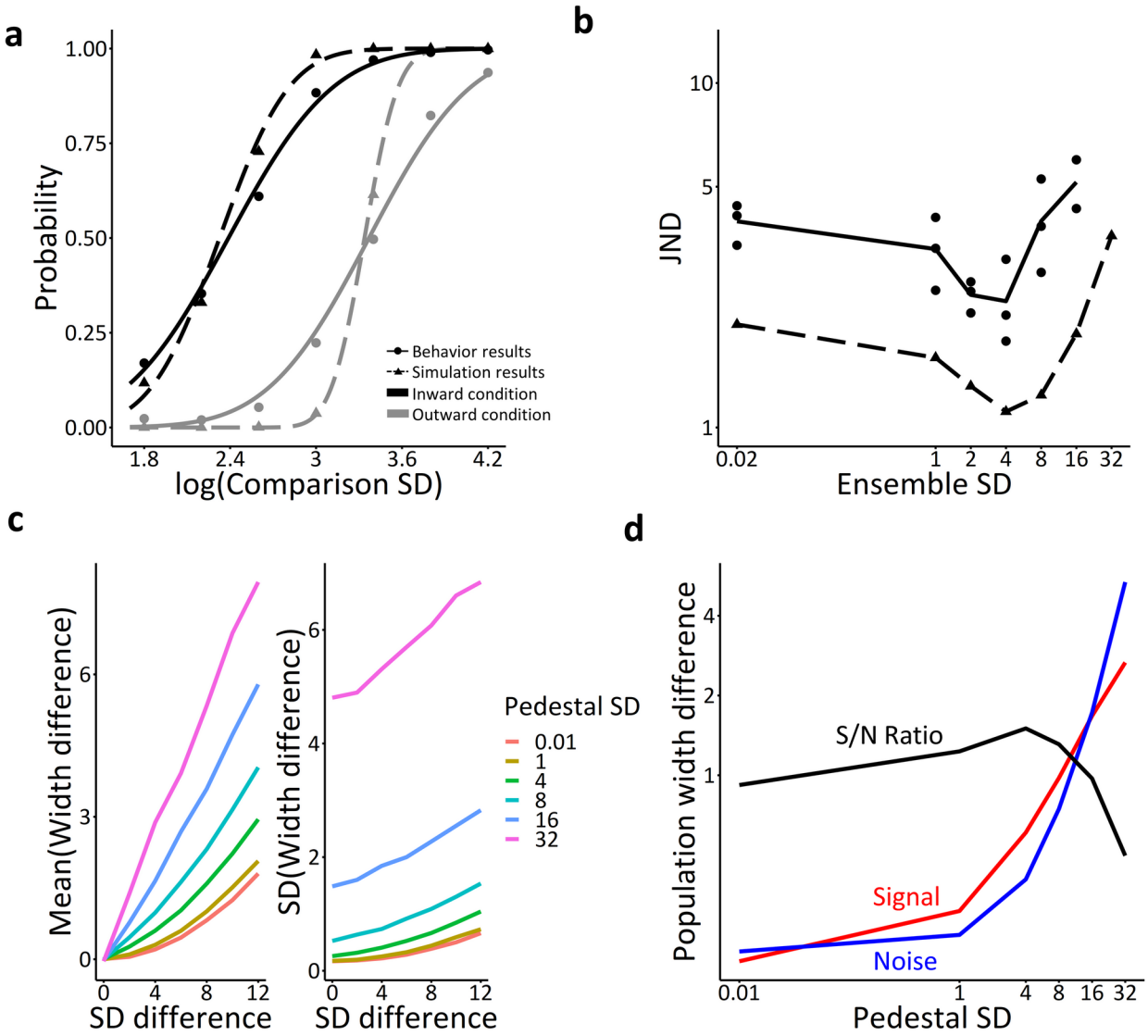


Figure 5. The results of orientation variance discrimination. (a) The behavioral and modeling results of Jeong & Chong (2021). Black lines indicate the inward condition and gray lines indicate outward condition. (b) The behavioral and modeling results of Morgan et al. (2008). The solid lines indicate behavioral results, and the dotted lines indicate modeling results. (c) The left panel shows how the signal strength varies depending on stimuli variance and the right panel shows how the noise changes depending on stimuli variance. (d) The signal-to-noise ratio of variance discrimination initially increases with stimuli variance and then decreases after a certain variance because of the different growth rate between the signal and noise.

Our model predicts that standard deviation is a better descriptor for people's variance discrimination (Jeong & Chong, 2021), because standard deviation can be easily estimated by model population responses. Figure 6 depicts a psychometric function and the derivation of a JND simulated by our model that generates two population responses for each of the stimuli and

compares their output bandwidths (SD's) as proxies for variances. Figure 5a shows that psychometric functions from the model successfully simulated the inward and outward conditions from Jeong and Chong (2021). The model also explains the variance discrimination pattern as a function of physical pedestal variance (including the dip) without introducing an additional assumption of thresholding internal noise, as in Morgan et al. (2008). Figure 5b shows that the overall shape resembles those reported by Morgan et al. (2008), with a smooth decline before the dip and a steady increment after the dip. Figures 5c and 5d explain the emergence of the dip by the difference in relative growth of signal strength and noise in variance representation as a function of stimulus variance. It shows that the signal strength (mean bandwidth difference between two population responses) grows faster than the noise (SD of bandwidth differences) up to some point causing the signal-to-noise ratio to improve and sharpen JND's. After that point, the signal strength grows slower than the noise and, thus, the signal-to-noise ratio steadily decreases. Therefore, our model suggests a very simple mechanistic explanation of the seemingly complex, non-monotonic pattern of variance discrimination without any additional assumption. Figure 5b plots behavioral (Morgan et al., 2008) and model JNDs against ensemble set's SD based on the model. The correlation between the behavioral and population JNDs was $r = .71$.

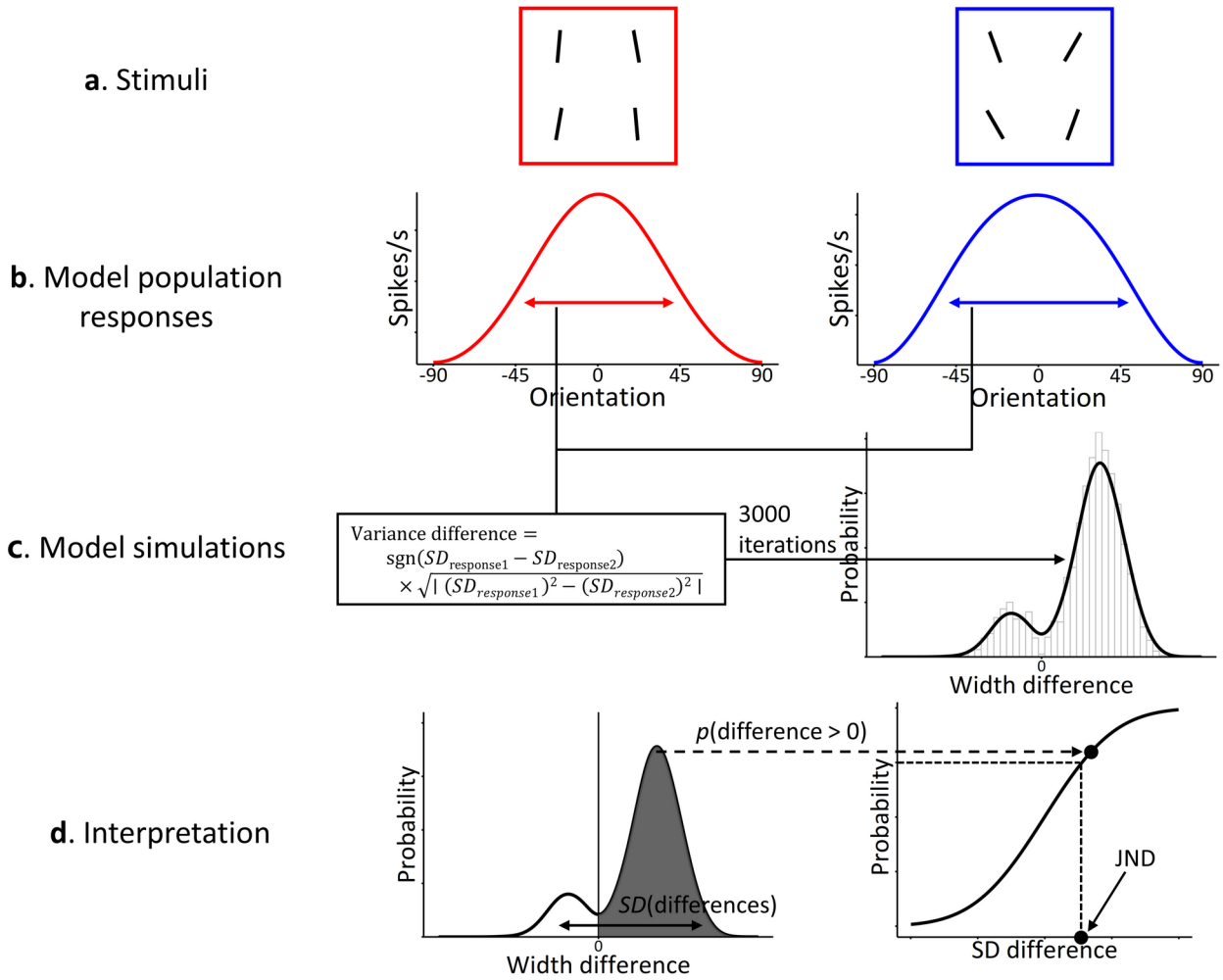


Figure 6. The procedure of orientation variance discrimination by the model. (a) Two sets of oriented bars. The standard deviation of the left set was 5.8 and that of the right set was 11.5. (b) Model population responses to each set of stimuli. (c) We simulated population response differences for 3,000 times to make their distribution. (d) Based on these distributions, we constructed a psychometric function and reads a JND from it.

Feature distribution learning. Up to this point, we applied the population coding model to various tasks requiring observers to estimate ensemble summary statistics, such as average or variance. Although these summary statistics convey the important information about all presented features, they are inevitably compressed and fail to provide more nuanced information about the representation of the whole feature distribution. However, recent research using a feature distribution learning (FDL) paradigm has shown that the visual system is able to implicitly reproduce the shape of a feature distribution, sensitive to local changes of the probability density (Chetverikov et al., 2016, 2017), which cannot be reduced to mean and variance. In the FDL task, participants usually perform an odd-one-out visual search task, that is, they look for a target with a unique feature that can change unpredictably from trial to trial. Critically, the trial sequence in the FDL task is organized in streaks (e.g., 5-6 trials in a row) where distractor features are drawn from exactly the same feature distribution and switches when

the distractor distribution shifts along a feature dimension and the target takes one of the values from the former distractor distribution. This allows to probe aftereffects of the previous feature distribution. Although only one point at this distribution can be probed in each streak switch, the whole distribution can be eventually constructed from multiple switches experiment-wide. One of the principal findings in the FDL research was the inhibitory effect of the former distractor distribution on the search time in switch trials. Most importantly, the amount of inhibition was proportional to the probability density of the probed feature value in the former distribution. Based on that finding, Chetverikov and colleagues (2016, 2017) concluded that the whole probability density function can be represented, yet mostly implicitly (Hansmann-Roth et al., 2021; but see Oriet & Hozempa, 2016).

Our model of ensemble population coding implies that the probabilistic feature distribution should be encoded along with summary statistics in a straightforward way. Since pooling neurons respond to input features in accordance with their tuning preferences, it is natural that more frequent features will elicit a stronger pooled response in a subpopulation of neurons tuned to these features. Thus, the overall shape of the population response necessarily reflects the physical feature distribution, as does the reaction time in the FDL paradigm. We argue that the RT functions in the FDL paradigm reflect ensemble representations in a pooled population response. This claim is based on quantitative demonstrations of strong resemblance (correlation) between the RT functions reported by Chetverikov and colleagues (2016) and population responses provided by our model. Critically, we show that this resemblance is considerably stronger than the resemblance between the RT functions and the physical feature distributions.

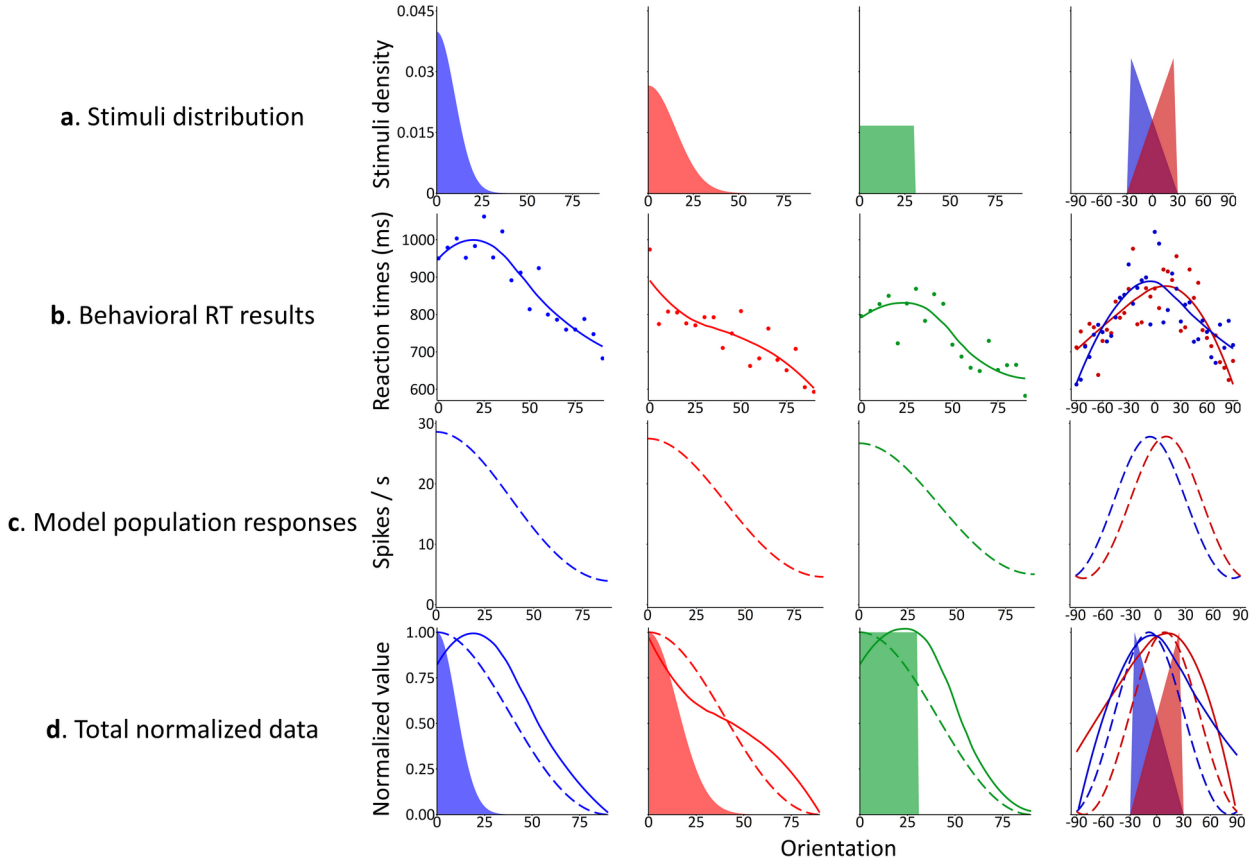


Figure 7. Feature distribution learning. (a) Various feature distributions used in Chetverikov et al. (2016). (b) RT results from Chetverikov et al. (2016). (c) Model population responses to each feature distribution. (d) Feature distributions, RT results, and model population responses together.

In their studies, Chetverikov and colleagues (2016, 2017) demonstrated distribution learning in orientation and color domains. The most common comparison made in these studies to demonstrate a specific effect of the distribution shape on the RT is a comparison between a Gaussian distribution and a uniform distribution (Figure 7a). Another sensitive case tested by this group is a comparison between asymmetric, triangular distributions with positive vs. negative skewness. We modelled both these cases, as they are most relevant to and illustrative of the representation of feature distribution shapes. Using the same early noise and tuning SD parameters for orientation defined in the Average discrimination in a 2-AFC task section, we found that the RT as a function of a probed orientation correlated with spike rates of model orientation-selective neurons within a range of r = .75 ~ .88, which was greater than the correlations between the RT and real probability densities (r = .43 ~ .69): e.g., r = .88 (RT-model spike rate correlation) vs. .43 (RT-real probability density correlation) in a Gaussian distribution with SD = 10 deg., r = .79 vs. .69 in a Gaussian distribution with SD = 15, r = .80 vs. .58 in a uniform distribution, r = .79 vs. .55 in a positively skewed triangular distribution, r = .75 vs. .60 in a negatively skewed triangular distribution. Note that the correlations between the RT data and the model spike rates were only slightly lower than the correlations between the RT data and

regression fits to these data (r = .79 ~ .93) provided by Chetverikov et al. (2016). Figure 7 shows visualizations of (a) stimuli distributions, (b) behavioral RT results, (c) model population response shapes, and (d) all three of them together that give intuitive understanding of why the population response is correlated with the RT stronger than with the actual probability density function. One strong driver of this is that the RT distributions and population responses both have extended smooth tails that the real distributions do not have and that cover a substantial range of the feature space. Second, the population response smoothens the difference between the shapes of various distributions and increases the representation at the center. This tendency is especially obvious in the comparison of two differently skewed triangular distributions where the differences between peak locations in the data and in the population codes (24 and 18 degrees, respectively) are much smaller than the difference (50 degrees) between peaks of the real distributions. This smoothing tendency also works for the uniform and Gaussian distributions (Figure 7), although our model exaggerates the resulting similarity of the corresponding population responses compared to the data. Both the extended tails and the smoothened contrasts between the distributions are explained by the way widely tuned neurons respond to individual features and then pooled together (see The architecture of the Core model section for explanation).

Discussion

The main goal of the current study was to provide a unified computational framework of ensemble perception that could account for the variety of separate findings in this field and to test it against accrued data from numerous experiments, tasks, and domains. We found that the single model of pooling and population coding captures the observed data. Specifically, the model reproduces performance in averaging tasks including most important effects such as the variability and set size effects. Furthermore, the model showed its ability to account for variance perception including the non-monotonic relationship between stimulus variance and the precision of variance estimation (the 'dipper' function, Morgan et al., 2008). Since the pooled population code is a distribution of neural responses depending on a feature distribution in the stimulus, the shape of the population response in our model conveys information about the whole distribution in a natural way. This provides an account for the priming effects of the feature distribution on subsequent responses to features from that distribution (Chetverikov et al., 2017). Finally, our model provides a qualitative explanation for adaptation aftereffects of ensembles and clustering effects of similar items, which we will discuss below.

It is important to note that the model rests on the same set of assumptions and only two fixed parameters (early noise and the width of the neuron's tuning curve) in an attempt to explain the results of previous studies across various domains and tasks, without introducing extra ad-hoc assumptions or parameters to account for a specific effect. That is, the model gives a parsimonious explanation of multiple phenomena with the same set of mechanisms. In this sense, the model opposes the most elaborated quantitative models of ensemble perception that are based on the idea of limited-capacity, noisy arithmetic computations (sampling models, Allik et al., 2013; Dakin, 2001; Solomon, 2010; Solomon et al. 2011). As an illustrative example, evidence has been found in some studies that, when observers judge a mean feature, they give more weight to items from the middle of the feature distribution and devalue highly deviant features (De Gardelle & Summerfield, 2011; Li et al., 2017), which in theory makes averaging robust against the influence of unusual items and outliers (Haberman & Whitney, 2010; Epstein et al., 2020). The existing models of this effect suggest various versions of a sampling strategy where some of the individual items literally gain different weights prior to averaging. For example, weights can be assigned by the likelihood estimation of each feature given the entire feature distribution. It is easy to see that this assumption introduces a whole new computational mechanism of likelihood estimation. More importantly, this mechanism is circular or at least very complicated: It implies that the whole distribution should be known before the items are weighted and that sampling and weighting are further needed to compute the mean. In contrast, in our model robust averaging is an inherent result of the way that individual signals are pooled and it requires no additional sampling and unequal weighting. The "machinery" of robust averaging is most clearly shown in our "toy" model (Figure 1b), where middle neurons get more activation because they accumulate signals from both sides of the feature distribution.

Another example of a problematic assumption of the sampling models is a claim that the efficient sample size slowly grows with the actual set size with the rate of the square root of all presented items (Dakin, 2001; Whitney & Yamanashi Leib, 2018). This was found based on modelling and the review of previous results, but its mechanism was not introduced. It is a serious theoretical question as to how the visual system increases its capacity as a function of set size or why it samples only 2 items in set size 4 when it is able to sample 4 items in set size 16. The effective sample size that is an important parameter of many ensemble models may not in fact reflect a real psychological construct. A recent work in the field of visual working memory (Shurgin et al., 2020) suggests a solid demonstration of how previous influential models used a default assumption of fixed memory capacity for discrete items (which is roughly a synonym for limited sample size in the ensemble models) and turned this assumption to a quantifiable model parameter, although this capacity limit could be not real (which is also supported by plausible

neurophysiological mechanisms - Bays, 2015). The pooling mechanism presented in our model does not have the sample size parameter and takes into account everything that falls into the receptive field of the pooling population of neurons.

Importantly, our model portrays the computation of ensemble statistics as an inherent result of feedforward processing in the visual system. The mechanisms of pooling and population coding are well established in the previous research. This implies that there is no need to postulate any special "statistical processor" or any separate computational mechanism to process ensemble statistics. Rather, ensemble statistical representations can arise in the same populations of neurons that are tuned to respond to various sensory properties of stimuli. The way these neural populations represent ensemble statistics, as our model suggests, has little to do with regular mathematical computation of statistics. For example, the average ensemble feature is not accomplished by summing up individual sampled features and dividing them by a sample size, as arithmetic averaging implies. Instead, our model suggests that the average feature is accomplished more directly (without the need to represent sums and the quantities as intermediate computational steps) via the peak activation of pooled responses. Again, this sort of computation is neurally plausible, as it is in line with well established mechanisms of competitive feature interactions in large receptive fields without focused attention (Desimone & Duncan, 1995; Kastner et al., 1998; Maunsell, 2015). This is an important caveat for many quantitative models of ensemble perception that literally apply the arithmetic way of calculating summary statistics to their simulations of observer's behavior in ensemble tasks (Allik et al., 2013; Baek & Chong, 2020a; Dakin, 2001; Solomon, 2010). Although the use of arithmetic rules is based on the strict definition of what summary statistics are, it can nevertheless be misleading in terms of how it is accomplished by the visual system.

Since our model implies ensemble coding by neurons tuned specifically to visual features, it can potentially account for a set of adaptation aftereffects of ensemble statistics (Corbett et al., 2012; Jeong & Chong, 2020; Maule & Franklin, 2020; Norman et al., 2015). Adaptation aftereffects of various features are often taken as evidence for the direct coding of these features by sensory systems (Webster, 2011, 2015). Importantly, the aftereffects are usually well explained by changes in population responsiveness induced by adaptation. Since our model also has population responses in its core, it has potentially a straightforward mechanism of the ensemble aftereffects. For example, when observers are adapted to an ensemble with a large mean size and then are shown an ensemble with a smaller mean size, the peak of the population response to the latter will be shifted to the left (towards even smaller size) because the right 'tail' of the population response distribution overlaps with the adapted population and, hence, attenuated. This will shift the estimated mean to a smaller size - this is exactly what is observed

in the data (Corbett et al., 2012). There are also adaptation aftereffects of variability that are more complicated (e.g., Jeong & Chong, 2020; Norman et al., 2015) but our model is in principle applicable to them as well, because the model also suggests the direct encoding of variability in population responses. Modelling the adaptation aftereffects of ensemble statistics is one of the future directions of developing our model.

In our model, population responses are more tightly formed with similar items. When there are two groups with sufficiently different across-group features and similar within-group features (e.g., yellowish flowers and greenish leaves), our model will form two distinct population responses with two different peaks and thereby distinguish two groups of features naturally without assuming additional mechanism of grouping (as shown in Figure 2a, bottom plot). This characteristic of our model can explain the difficulty of visual search, when distractors form multiple groups (Utochkin & Yurevich, 2016). It can also explain similarity-based clusters in visual working memory (Son et al., 2020). Modelling the clustering effect of ensembles is another future direction.

The idea of representing ensemble statistics via pooling in large receptive fields has repercussions in broader theories of visual perception. Ensemble statistics can be thought of as a form of gist representation (Cohen et al., 2016), that is, the rough impression of the whole scene without knowing details. A prominent 'reverse hierarchy theory' (Ahissar & Hochstein, 2002; Hochstein & Ahissar, 2004) links conscious perception to feedforward and feedback processing streams in the visual system. The theory suggests that the gist percept arises at the top of the fast feedforward stream, where neurons with large receptive fields respond to large portions of the visual field; whereas detailed vision requires slow feedback propagation of focused attention signals to lower-level neurons with small receptive fields. In our model, ensemble representations also emerge at neurons with large receptive fields as a result of feedforward pooling. In line with the reverse hierarchy theory, this predicts that conscious ensemble perception precedes the recognition of individual items. This prediction was recently supported neurophysiologically (Epstein & Emmanouil, 2021).

Although our model in its current simple form accounts for many basic patterns of ensemble perception, it can be further augmented with additional mechanisms to take into account other effects not covered in the current work. For example, although the current version of the model explains the core mechanism of building ensemble representations in a feedforward stream of processing, it does not consider modulatory effects of focused attention (Choi & Chong, 2020; De Fockert & Marchant, 2008; Kanaya et al., 2018; Iakovlev & Utochkin, 2021) and working memory (Williams et al., 2021) that can bias ensemble statistical estimates toward a subset of items. In our model, this can be accomplished by giving unequal weights to attended

and unattended item representations when they enter from the bottom layer to the top, pooling layer. Alternatively, tuning curves of pooling neurons can be directly scaled up for attended features or scaled down for unattended features, which would reflect a neurally plausible mechanism of the biasing role of attention in neural representations within large receptive fields (Desimone & Duncan, 1995; Kastner et al., 1998; Maunsell, 2015). In addition, feedback connections from the top layer to the bottom layer are also possible. These connections can provide the modulation of low-level neural responses given the population response on top. The feedback modulation was earlier suggested as a potential mechanism of dealing with outliers (Haberman & Whitney, 2012) that allows the visual system to detect odd-one-out elements in the ensemble efficiently as a function of their difference from the rest of items (Hochstein et al., 2018; Rosenholtz, 2001). On the other hand, this feedback modulation can reduce the influence of outliers on the overall representation of an ensemble (Haberman & Whitney, 2010) in a slow, iterative loop process (Epstein et al., 2020).

Despite its simplicity and formal appropriateness to account for ensemble perception in all domains, the model has theoretical limitations. Most substantially, our spatial pooling mechanism implies two constraints based on neural plausibility. First, there should be an ensemble (or a texture) of spatially dispersed elements in order to represent individual elements in independent receptive fields and then pool them by populations with larger receptive fields. Therefore, we do not assume that there is exactly the same mechanism for averaging temporal sequences presented in the same receptive fields (unless a temporal integrator is assumed that can pool over time). Second, the model should work for the features that can be initially processed at a level with relatively small receptive fields, so that there should be a layer with larger receptive fields that can pool from that former one. It can be problematic for complex objects (e.g., multiple faces with various expressions) and scene features that are detected at very high levels of the visual cortex with very large receptive fields. One way to deal with this issue is assuming that these features can be reducible to more basic features detectable at relatively lower levels or that these basic features are represented in multivariate distributions. Alternatively, ensemble representations of high-level features can involve different mechanisms that have little overlap with low-level ensemble representations (Haberman et al., 2015).

Concluding remarks

Ensemble representations 1) include the average information more than constituents, 2) consist of various statistical properties such as variance and distributional properties, 3) are formed over many stages of visual processing, 4) become precise if included items are similar

and more items are included, and 5) are useful for many visual functions. To provide a theoretical and computational framework for these various facets of ensemble perception, we proposed a population coding model of ensemble perception. It consists of a simple feature layer and a pooling layer. We assumed ensemble representations as population responses in the pooling layer and decoded various statistical summaries from population responses. Our model successfully predicted averaging performance in orientation, size, color, and direction across different tasks. Furthermore, it predicted variance discrimination performance and the priming effects of feature distributions. Finally, it explained the well-known variance and set size effects and has a potential for explaining the adaptation, clustering, and robust averaging effects.

References

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences, 8*(10), 457-464.

Albright, T. D. (1984). Direction and orientation selectivity of neurons in visual area MT of the macaque. *Journal of Neurophysiology*, *52*(6), 1106-1130.

Allik, J., Toom, M., Raidvee, A., Averin, K., & Kreegipuu, K. (2013). An almost general theory of mean size perception. *Vision Research, 83*, 25-39.

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

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science, 19*(4), 392-398. http://doi.org/10.1111/j.1467-9280.2008.02098.x

Ariely, D. (2001). Seeing sets: representation by statistical properties. *Psychological Science, 12*(2), 157-162. http://doi.org/10.1111/1467-9280.00327

Baek, J., & Chong, S. C. (2020a). Distributed attention model of perceptual averaging. *Attention, Perception, & Psychophysics, 82*(1), 63-79.

Baek, J., & Chong, S. C. (2020b). Ensemble perception and focused attention: Two different modes of visual processing to cope with limited capacity. *Psychonomic Bulletin & Review, 27*(4), 602-606.

Bauer, B. (2009). Does Stevens's Power Law for Brightness Extend to Perceptual Brightness Averaging? *Psychological Record, 59*, 171-185.

Bays, P. M. (2015). Spikes not slots: noise in neural populations limits working memory. *Trends in Cognitive Sciences*, *19*(8), 431-438.

Berens, P. (2009). CircStat: A MATLAB toolbox for circular statistics. *Journal of Statistical Software, 31*, 1-21.

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: ensemble statistics bias memory for individual items. *Psychological Science, 22*(3), 384–392. http://doi.org/10.1177/0956797610397956

Boussaoud, D., Desimone, R., & Ungerleider, L. G. (1991). Visual topography of area TEO in the macaque. *Journal of Comparative Neurology, 306*(4), 554-575.

Brezis, N., Bronfman, Z. Z., & Usher, M. (2018). A Perceptual-Like Population-Coding Mechanism of Approximate Numerical Averaging. *Neural Computation, 30*(2), 428-446.

Broadbent, D. E. (1958). *Perception and communication*. London: Pergamon Press.

Brouwer, G. J., & Heeger, D. J. (2009). Decoding and reconstructing color from responses in human visual cortex. *Journal of Neuroscience, 29*(44), 13992-14003.

Cavanagh, P. (2001). Seeing the forest but not the trees. *Nature Neuroscience, 4*(7), 673-674.

Cha, O., & Chong, S. C. (2018). Perceived average orientation reflects effective gist of the surface. *Psychological Science, 29*(3), 319-327.

Chetverikov, A., Campana, G., & Kristjánsson, Á. (2016). Building ensemble representations: How the shape of preceding distractor distributions affects visual search. *Cognition*, *153*, 196-210.

Chetverikov, A., Campana, G., & Kristjánsson, A. (2017). Representing color ensembles. *Psychological Science*, *28*(10), 1510-1517. DOI:10.1177/0956797617713787

Choi, Y. M., & Chong, S. C. (2020). Effects of selective attention on mean-size computation: Weighted averaging and perceptual enlargement. *Psychological Science, 31*(10), 1261-1271.

Chong, S. C., Joo, S. J., Emmmanouil, T. A., & Treisman, A. (2008). Statistical processing: Not so implausible after all. *Perception & Psychophysics, 70*(7), 1327-1334.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research, 43*, 393-404. http://doi.org/10.1016/S0042-6989(02)00596-5

Cohen, M. A., Dennett, D. C., & Kanwisher, N. (2016). What is the bandwidth of perceptual experience? *Trends in Cognitive Sciences, 20*, 324-335. DOI:10.1016/j.tics.2016.03.006

Corbett, J. E. (2017). The whole warps the sum of its parts: Gestalt-defined-group mean size biases memory for individual objects. *Psychological Science*, *28*(1), 12-22.

Corbett, J. E., & Melcher, D. (2014). Stable statistical representations facilitate visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(5), 1915-1925.

Corbett, J. E., Wurnitsch, N., Schwartz, A., & Whitney, D. (2012). An aftereffect of adaptation to mean size. *Visual Cognition, 20*(2), 211-231.

Dakin, S. C. (2001). Information limit on the spatial integration of local orientation signals. *Journal of the Optical Society of America A, 18*(5), 1016-1026.

Dakin, S. C. & Watt, R. J. (1997). The computation of orientation statistics from visual texture. *Vision Research, 37,* 3181-3192.

De Fockert, J. W., & Marchant, A. P. (2008). Attention modulates set representation by statistical properties. *Perception & Psychophysics*, *70*(5), 789-794.

De Gardelle, V., & Summerfield, C. (2011). Robust averaging during perceptual judgment. *Proceedings of the National Academy of Sciences*, *108*(32), 13341-13346.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18*(1), 193-222.

Desimone, R., & Schein, S. J. (1987). Visual properties of neurons in area V4 of the macaque: sensitivity to stimulus form. *Journal of Neurophysiology*, *57*(3), 835-868.

Duffy, S., Huttenlocher, J., Hedges, L. V., & Crawford, L. E. (2010). Category effects on stimulus estimation: Shifting and skewed frequency distributions. *Psychonomic Bulletin & Review*, *17*(2), 224-230. https://doi.org/10.3758/PBR.17.2.224

Ester, E. F., Serences, J. T., & Awh, E. (2009). Spatially global representations in human primary visual cortex during working memory maintenance. *Journal of Neuroscience*, *29*(48), 15258-15265.

Epstein, M. L., & Emmanouil, T. A. (2021). Ensemble Statistics Can Be Available before Individual Item Properties: Electroencephalography Evidence Using the Oddball Paradigm. *Journal of Cognitive Neuroscience*, *33*(6), 1056-1068.

Epstein, M. L., Quilty-Dunn, J., Mandelbaum, E., & Emmanouil, T. A. (2020). The outlier paradox: The role of iterative ensemble coding in discounting outliers. *Journal of Experimental Psychology: Human Perception and Performance, 46*(11), 1267-1279. https://doi.org/10.1037/xhp0000857

Galton, F. (1907). Vox populi (the wisdom of crowds). *Nature, 75*(7), 450-451.

Georgopoulos, A. P., Schwartz, A. B., & Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science, 233*(4771), 1416-1419.

Haberman, J., Brady, T. F., & Alvarez, G. A. (2015). Individual differences in ensemble perception reveal multiple, independent levels of ensemble representation. *Journal of Experimental Psychology: General, 144*(2), 432–446. https://doi.org/10.1037/xge0000053

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology, 17*(17), R751-R753. http://doi.org/10.1016/j.cub.2007.06.039

Haberman, J. & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention, Perception, & Psychophysics, 72*, 1825-1838.

Haberman, J., & Whitney, D. (2012). Ensemble perception: Summarizing the scene and broadening the limits of visual processing. In J. Wolfe and L. Robertson (Eds.), *From Perception to Consciousness: Searching with Anne Treisman.* Oxford University Press, 339-349.

Hansmann-Roth, S., Kristjánsson, Á., Whitney, D., & Chetverikov, A. (2021). Dissociating implicit and explicit ensemble representations reveals the limits of visual perception and the richness of behavior. *Scientific Reports, 11*, 1-12.

Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron, 36*(5), 791-804.

Hochstein, S. Pavlovskaya, M., Bonneh, Y. & Soroker, N. (2018). Comparing set summary statistics and outlier pop out in vision. *Journal of Vision, 18*(13):12, 1-13.

Iakovlev, A. U., & Utochkin, I. S. (2021). Roles of saliency and set size in ensemble averaging. *Attention, Perception, & Psychophysics, 83*(3), 1251-1262.

Im, H. Y., & Halberda, J. (2013). The effects of sampling and internal noise on the representation of ensemble average size. *Attention, Perception & Psychophysics, 75*(2), 278-286. http://doi.org/10.3758/s13414-012-0399-4

Jeong, J., & Chong, S. C. (2020). Adaptation to mean and variance: Interrelationships between mean and variance representations in orientation perception. *Vision Research, 167*, 46-53.

Jeong, J., & Chong, S. C. (2021). Perceived variability reflects the reliability of individual items. *Vision Research, 183*, 91-105.

Kastner, S., De Weerd, P., Desimone, R., & Ungerleider, L. G. (1998). Mechanisms of directed attention in the human extrastriate cortex as revealed by functional MRI. *Science, 282*(5386), 108-111.

Kanaya, S., Hayashi, M. J., & Whitney, D. (2018). Exaggerated groups: Amplification in ensemble coding of temporal and spatial features. *Proceedings of the Royal Society B: Biological Sciences, 285*(1879), 20172770.

Khayat, N. & Hochstein, S. (2018). Perceiving set mean and range: Automaticity and precision. *Journal of Vision, 18*(9):23, 1-14.

Khayat, N. & Hochstein, S. (2019). Relating categorization to set summary statistics perception. *Attention, Perception & Psychophysics, 75*(2), 278-286.

Khayat, N., Fusi, S., & Hochstein, S. (2021). Perceiving ensemble statistics of novel image sets. *Attention, Perception, & Psychophysics, 83*(3), 1312-1328.

Khvostov, V. A., & Utochkin, I. S. (2019). Independent and parallel visual processing of ensemble statistics: Evidence from dual tasks. *Journal of Vision, 19*(9):3, 1-18.

Kim, M., & Chong, S. C. (2020). The visual system does not compute a single mean but summarizes a distribution. *Journal of Experimental Psychology: Human Perception and Performance, 46*(9), 1013-1028. doi:http://dx.doi.org/10.1037/xhp0000804

Lee, H., Baek, J., & Chong, S. C. (2016). Perceived magnitude of visual displays: Area, numerosity, and mean size. *Journal of Vision, 16*(3):12, 1-11.

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*(8), e1005723.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*, 279-281. DOI:10.1038/36846

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology Press.

Manassi, M., Liberman, A., Chaney, W., & Whitney, D. (2017). The perceived stability of scenes: serial dependence in ensemble representations. *Scientific Reports, 7*, 1-9.

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. http://doi.org/10.1016/j.actpsy.2012.11.002

Maule, J., & Franklin, A. (2015). Effects of ensemble complexity and perceptual similarity on rapid averaging of hue. *Journal of Vision, 15*(4):6, 1-18. http://doi.org/10.1167/15.4.6

Maule, J., & Franklin, A. (2016). Accurate rapid averaging of multihue ensembles is due to a limited capacity subsampling mechanism. *Journal of the Optical Society of America A: Optics, Image Science, and Vision, 33*, A22– A29. doi:10.1364/JOSAA.33.000A22

Maule, J., & Franklin, A. (2020). Adaptation to variance generalizes across visual domains. *Journal of Experimental Psychology: General*, *149*(4), 662-675. doi:http://dx.doi.org/10.1037/xge0000678.

Maunsell, J. H. (2015). Neuronal mechanisms of visual attention. *Annual Review of Vision Science, 1*, 373-391.

Maunsell, J. H., & Van Essen, D. C. (1983). Functional properties of neurons in middle temporal visual area of the macaque monkey. I. Selectivity for stimulus direction, speed, and orientation. *Journal of Neurophysiology, 49*(5), 1127-1147.

McAdams, C. J., & Maunsell, J. H. (1999). Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. *Journal of Neuroscience, 19*(1), 431-441.

Morgan, M., Chubb, C., & Solomon, J. A. (2008). A 'dipper' function for texture discrimination based on orientation variance. *Journal of Vision, 8*(11):9, 1-8 .

Motoyoshi, I., Nishida, S. Y., Sharan, L., & Adelson, E. H. (2007). Image statistics and the perception of surface qualities. *Nature, 447*(7141), 206-209.

Oriet, C., & Hozempa, K. (2016). Incidental statistical summary representation over time. *Journal of Vision, 16*(3):3, 1-14.

Palmer, E.M., Fencsik, D. E., Flusberg, S. J., Horowitz, T. S., & Wolfe, J. M. (2011). Signal detection evidence for limited capacity in visual search. *Attention, Perception & Psychophysics, 73*(8), 2413-2424. https://doi.org/10.3758/s13414-011-0199-2

Norman, L. J., Heywood, C. A., & Kentridge, R. W. (2015). Direct encoding of orientation variance in the visual system. *Journal of Vision*, *15*(4):3, 1-14.

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. http://doi.org/10.1038/89532

Raidvee, A., Toom, M., & Allik, J. (2021). A method for detection of inattentional feature blindness. *Attention, Perception, & Psychophysics, 83*(3), 1282-1289.

Roberts, T., Cant, J. S., & Nestor, A. (2019). Elucidating the neural representation and the processing dynamics of face ensembles. *Journal of Neuroscience*, *39*(39), 7737-7747.

Rodman, H. R., & Albright, T. D. (1987). Coding of visual stimulus velocity in area MT of the macaque. *Vision Research*, *27*, 2035-2048.

Robitaille, N., & Harris, I. M. (2011). When more is less: Extraction of summary statistics benefits from larger sets. *Journal of Vision, 11*(12):18, 1-8. http://doi.org/10.1167/11.12.18

Rosenholtz, R. (2000). Significantly different textures: A computational model of pre-attentive texture segmentation. In Vernon D. (Ed.) *Proceedings of the European Conference on Computer Vision (ECCV '00), LNCS 1843*, 197-211. https://doi.org/10.1007/3-540-45053-X_13

Rosenholtz, R. (2001). Visual search for orientation among heterogeneous distractors: Experimental results and implications for signal-detection theory models of search. *Journal of Experimental Psychology: Human Perception and Performance, 27*(4), 985-999. DOI:10.1037//0096-1523.27.4.985

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.

Solomon, J. A. (2009). The history of dipper functions. *Attention, Perception, & Psychophysics, 71*(3), 435-443.

Solomon, J. A. (2010). Visual discrimination of orientation statistics in crowded and uncrowded arrays. *Journal of Vision*, *10*(14):19, 1-16.

Solomon, J. A., Morgan, M., & Chubb, C. (2011). Efficiencies for the statistics of size discrimination. *Journal of Vision*, *11*(12):13, 1-11.

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*(1), 46-59. doi:http://dx.doi.org/10.1037/xlm0000722

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31*(1), 137-149.

Sun, J., & Chong, S. C. (2020). Power of averaging: Noise reduction by ensemble coding of multiple faces. *Journal of Experimental Psychology: General, 149*(3), 550-563. doi:http://dx.doi.org/10.1037/xge0000667

Sweeny, T. D., Haroz, S., & Whitney, D. (2013). Perceiving group behavior: Sensitive ensemble coding mechanisms for biological motion of human crowds. *Journal of Experimental Psychology: Human Perception and Performance, 39*(2), 329-337.

Teghtsoonian, M. (1965). The judgment of size. *American Journal of Psychology, 78*(3), 392-402.

Treisman, A., & Gormican, S. (1988). Feature analysis in early vision: evidence from search asymmetries. *Psychological Review, 95*(1), 15-48.

Treue, S., Hol, K., & Rauber, H. J. (2000). Seeing multiple directions of motion – physiology and psychophysics. *Nature Neuroscience, 3,* 270-276.

Ungerleider, L. G., & Bell, A. H. (2011). Uncovering the visual "alphabet": advances in our understanding of object perception. *Vision Research, 51*, 782-799.

Utochkin, I. S. (2015). Ensemble summary statistics as a basis for rapid visual categorization. *Journal of Vision, 15*(4):8, 1-14. DOI:10.1167/15.4.8

Utochkin, I. S., & Brady, T. F. (2020). Individual representations in visual working memory inherit ensemble properties. *Journal of Experimental Psychology: Human Perception and Performance*, *46*(5), 458-473. doi:http://dx.doi.org/10.1037/xhp0000727

Utochkin, I. S., & Tiurina, N. A. (2014). Parallel averaging of size is possible but range-limited: A reply to Marchant, Simons, and De Fockert. *Acta Psychologica, 146*, 7-18. http://doi.org/10.1016/j.actpsy.2013.11.012

Utochkin, I. S., & Yurevich, M. A. (2016). Similarity and heterogeneity effects in visual search are mediated by "segmentability". *Journal of Experimental Psychology: Human Perception and Performance, 42*(7), 995-1007. DOI:10.1037/xhp0000203

Virtanen, L. S., Olkkonen, M., & Saarela, T. P. (2020). Color ensembles: Sampling and averaging spatial hue distributions. *Journal of Vision*, *20*(5):1, 1-14.

Watamaniuk, S. N., & Duchon, A. (1992). The human visual system averages speed information. *Vision Research, 32*, 931-941.

Watamaniuk, S. N. J., Sekuler, R., & Williams, D. W. (1989). Direction perception in complex dynamic displays: The integration of direction information. *Vision Research, 29*, 47-59.

Webster, M. A. (2011). Adaptation and visual coding. *Journal of Vision, 11*(5):3, 1-23.

Webster, M. A. (2015). Visual adaptation. *Annual Review of Vision Science, 1*, 547-567.

Williams, R. S., Pratt, J., Ferber, S., & Cant, J. S. (2021). Tuning the ensemble: Incidental skewing of the perceptual average through memory-driven selection. *Journal of*

Experimental Psychology: Human Perception and Performance, 47(5), 648-661. https://doi.org/10.1037/xhp0000907.

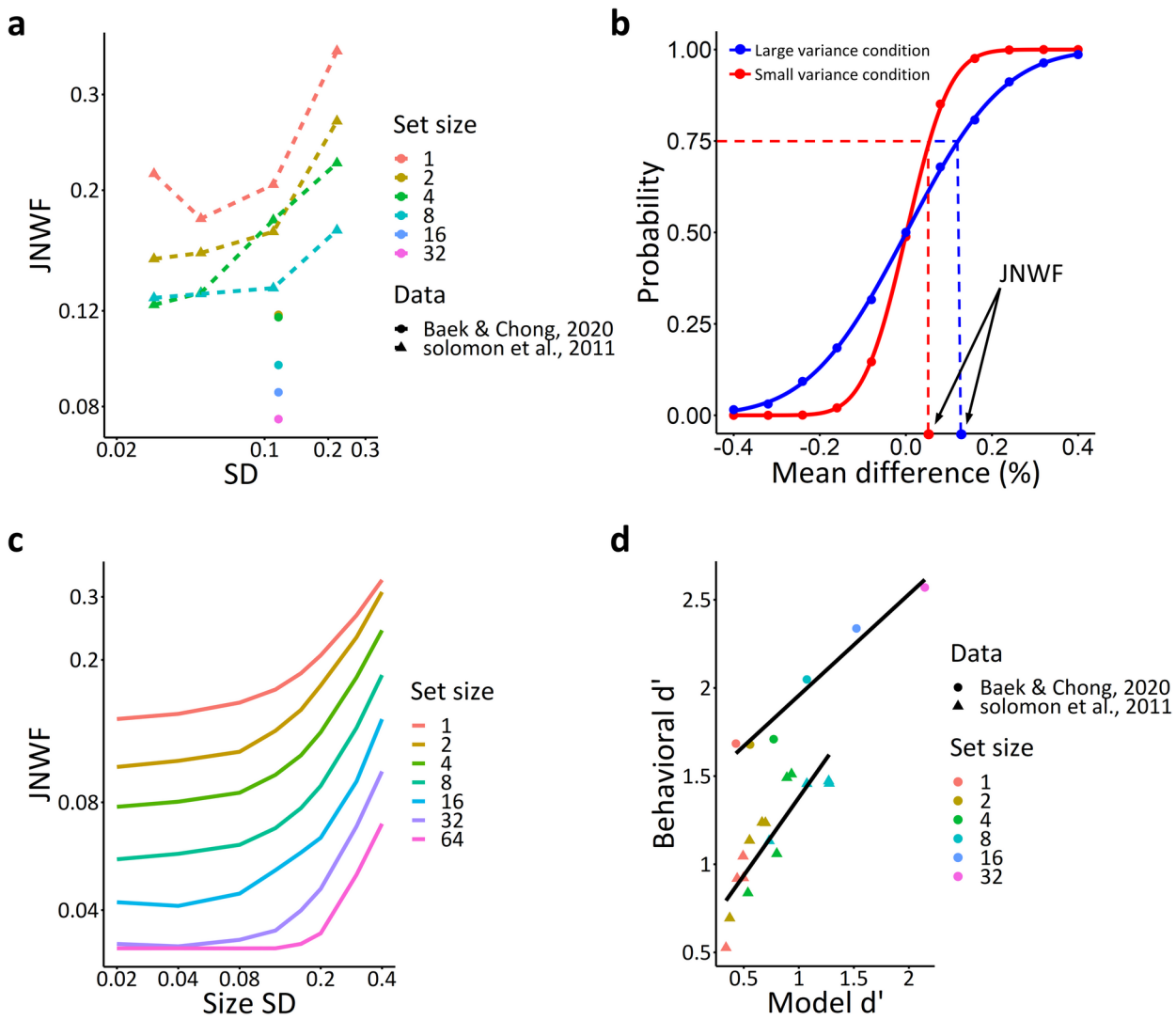
Whitney, D., & Yamanashi Leib, A. (2018). Ensemble perception. *Annual Review of Psychology, 69*, 105-129.

Yamanashi Leib, A., Chang, K., Xia, Y., Peng, A., & Whitney, D. (2020). Fleeting impressions of economic value via summary statistical representations. *Journal of Experimental Psychology: General, 149(*10), 1811-1822. doi:http://dx.doi.org/10.1037/xge0000745

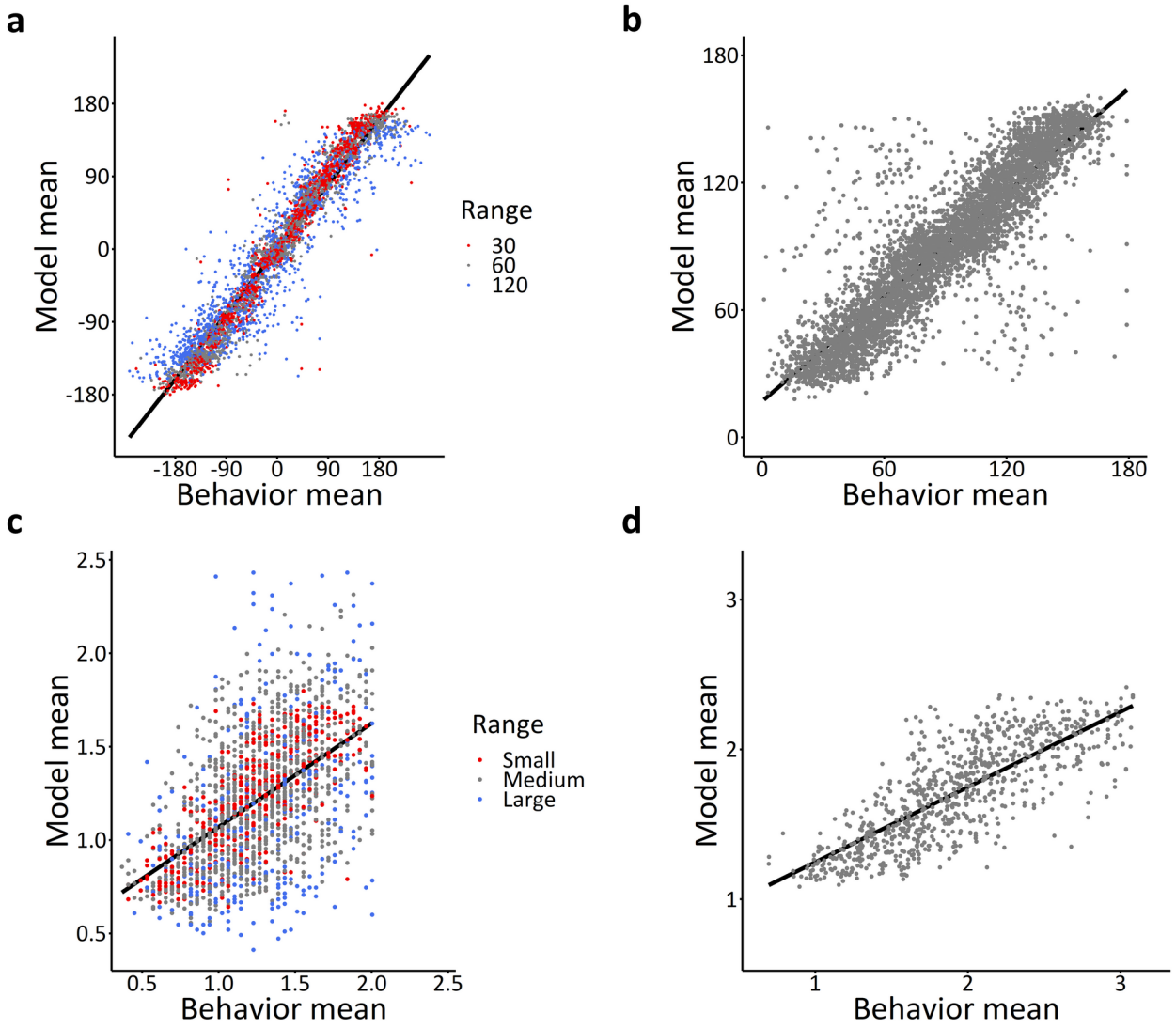
Yamanashi Leib, A., Kosovicheva, A. & Whitney, D. (2016). Fast ensemble representations for abstract visual impressions. *Nature Communications, 7*, 13186.

Yashiro, R., Sato, H., Oide, T., & Motoyoshi, I. (2020). Perception and decision mechanisms involved in average estimation of spatiotemporal ensembles. *Scientific Reports, 10*, 1-10.

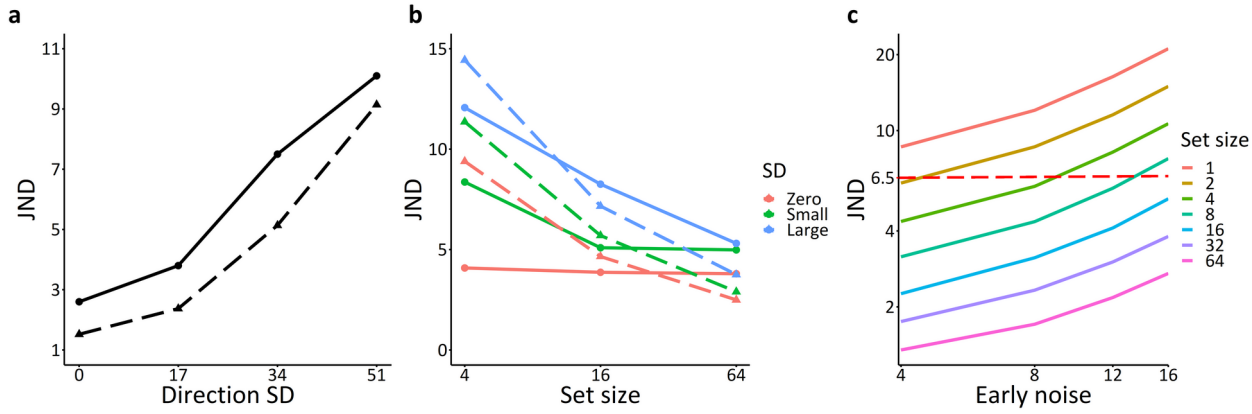
Supplementary figures



Supplementary Figure 1. The variance and set size effects of size averaging. (a) Previous studies have found that mean size discrimination performance decreased with increasing variance and decreasing set size (Baek & Chong, 2020; Solomon et al., 2011). (b) Psychometric functions of mean size discrimination. The blue and red lines indicate the results of the large and small variance conditions. (c) Model JNDs are plotted against SD and set size. They well capture the qualitative patterns of behavioral results presented in (a). (d) The correlation results between behavioral *d*' from previous studies (Baek & Chong, 2020; Solomon et al., 2011) and model *d*'.



Supplementary Figure 2. The results of an adjustment method. Orientation averaging results (a: Utochkin & Brady, 2020; b: Epstein et al., 2020) and size averaging results (c: Khvostov & Utochkin, 2019; d: Kim & Chong, 2020). Ranges in a is the same as in the original study. The range 1 in c indicates 0.3, the range 2 collapses 0.6, 0.9, and 1.2, and the range 3 indicates 1.5. We collapsed the some of conditions to visualize the variance effect better.



Supplementary Figure 3. (a) The simulation results of averaging motion directions (Watamaniuk et al., 1989). (b) The simulation results of color averaging (Virtanen et al., 2020). (c) Our model can explain no effect of set size on orientation averaging, if early noise increases with set sizes. Here, orientation discrimination JND of 6.5 was similar across different set sizes of 2, 4, 8, if we assume that early noise is increased with set sizes.