Ensemble perception of facial attractiveness

Anna X. Luo
Department of Psychiatry, University of British Columbia,
Vancouver, BC, Canada

Guomei Zhou
Department of Psychology, Sun Yat-sen University,
Guangzhou, Guangdong China

Ensemble perception, the extraction of a statistical summary of multiple instances of a feature, enables efficient processing of information. Here we investigated whether ensemble representations can be formed for facial attractiveness, a socially important complex feature. After verifying that our face stimuli produced by geometric morphing represented a valid continuum of attractiveness (Experiment 1), we asked participants to compare the average attractiveness of four faces with a single probe face. Whether the four faces were homogeneous or heterogeneous resulted in highly similar performance levels, suggesting the visual system could extract an ensemble representation of the attractiveness of a heterogeneous group of faces. Statistical simulations with human-level bias and noise indicated participants did not rely on subsampling one random face or the most/least attractive face from the array (Experiment 2). Ensemble perception of facial attractiveness was not affected by variance in the stimulus array (Experiment 3), did not depend on memory of individual faces in the array (Experiment 4), and could be extended to larger arrays with faces asymmetrically distributed around the set mean (Experiment 5). Our findings give further evidence to the prevalence of perception of statistical regularities in vision.

Introduction

The human visual system is found to be sensitive to various types of statistical properties embedded in real-world inputs. Not only is it adept in detecting first-order relationships of local sensors, such as the contrasts in luminance and orientation (Cataliotti & Gilchrist, 1995; Nothdurft, 1991), it is also sensitive to correlations of sensor response, such as the co-occurrence of local edges in natural images (Geisler, Perry, Super, & Gallogly, 2001) and correlations across orientations, positions, and scales in textures (Portilla & Simoncelli, 2000). Recent studies suggest another type of statistical perception from visual inputs: ensemble perception, the ability to extract the average value of a feature dimension from multiple objects and to use it as a compact representation of the complex environment (Alvarez, 2011). The ability of statistical averaging has been demonstrated with low-level feature domains, including size (Ariely, 2001; Chong & Treisman, 2003), orientation (Parkes, Lund, Angelucci, Solomon, & Morgan, 2001), brightness (Bauer, 2009), and location (Alvarez & Oliva, 2008). More recently, evidence shows that observers also form accurate ensemble representations for complex high-level features, such as facial expression (Haberman & Whitney, 2007; Haberman & Whitney, 2009), gender (Haberman & Whitney, 2007), and identity (J. de Fockert & Wolfenstein, 2009; J. W. de Fockert & Gautrey, 2013; Neumann, Schweinberger, & Burton, 2013). Importantly, ensemble perception can be formed without high-fidelity representation of individual objects in the group (Ariely, 2001; Haberman & Whitney, 2009) and with minimal requirement of focal attention (Alvarez & Olivia, 2008).

Group-level attractiveness perception

Ensemble perception has been demonstrated among a variety of simple and composite features, yet it remains unclear whether the visual system could extract ensemble representations of another high-level feature—facial attractiveness—from a heterogeneous group of faces. Importance of this question is twofold. First, physical attractiveness of faces is socially significant as it predicts mate choice (Grammer & Thornhill, 1994) and perceived intelligence, sociability, and altruism (Griffin & Langlois, 2006; Zebrowitz, Hall, Murphy, & Rhodes, 2002). If the relationship between facial attractiveness and perceived personal qualities or aptitude can be extended to social groups, then our ensemble perception of the overall facial attractiveness of a group might influence our judgment

Citation: Luo, A. X., & Zhou, G. (2018). Ensemble perception of facial attractiveness. *Journal of Vision*, *18*(8):7, 1–19, https://doi.org/10.1167/18.8.7.

https://doi.org/10.1167/18.8.7 Received June 28, 2017; published August 13, 2018 ISSN 1534-7362 Copyright 2018 The Authors

of group-level characteristics, forming a basis for social stereotypes and biases. Second, ensemble perception of facial attractiveness, in particular, the extraction of average attractiveness from multiple faces, could be an intermediate step for group-induced biases on attractiveness perception, including the cheerleader effect (Walker & Vul, 2014) and the group attractiveness effect (van Osch, Blanken, Meijs, & van Wolferen, 2015). In the cheerleader effect, individual faces are rated as more attractive when they are presented in photos of a group or when their photos are collocated with others' photos than when their photos are presented alone. Similarly, in the group attractiveness effect, observers judge the overall physical attractiveness of a group to be greater than the average rating of each member in isolation (van Osch et al., 2015). Authors of these two studies proposed perception of average attractiveness of the group as a potential explanation: Observers may unconsciously morph all the faces in the group into one average face, and this average face could bias their judgment of group attractiveness or group member attractiveness (van Osch et al., 2015; Walker & Vul, 2014). However, the possibility and efficiency of average attractiveness perception has yet to be explicitly examined and characterized.

Our study thus set out to systematically investigate whether perception of average facial attractiveness could occur for multiple faces. In van Osch et al. (2015) and Walker and Vul (2014), group-induced bias in attractiveness was derived from comparison of group (or group member) ratings and individual ratings, which may vary in perceptual load and processing time. This method is different from that of previous studies on ensemble coding of facial identity and emotion (J. de Fockert & Wolfenstein, 2009; J. W. de Fockert & Gautrey, 2013; Haberman & Whitney, 2007; Haberman & Whitney, 2009; Neumann et al., 2013), which is based on the comparison of discrimination thresholds between a homogeneous set and a heterogeneous set of faces. For example, in Haberman and Whitney (2007) and Haberman and Whitney (2009), which examined average facial expression computation, a set of identical faces (the homogeneous condition) or different faces (the heterogeneous condition) and a subsequent probe were presented to participants, and the participants' task was to judge whether the emotionality of the probe was greater than the average emotionality of the set. Ensemble coding was defined as that participants had equivalent performance in the homogeneous and the heterogeneous conditions. Comparing perceptual efficacy of homogeneous and heterogeneous sets allows holding the overall perceptual load constant. We thus employed this classic paradigm to examine perception of average facial attractiveness.

Ruling out heuristic strategies

One outstanding question regarding ensemble perception is whether all items in the group participate in and contribute equally to mean computation. Critics of ensemble perception point out that observers might employ heuristic strategies, rather than averaging, to derive an accurate response in tasks that purport to test mean computation. Myczek and Simons (2008) used noise-free ideal observer models to simulate mean size computation from an array of circles. They claim that subsampling one or two items from the array would be sufficient for the model to respond as accurately as human observers in a task of comparing the size of a probe and the average size of all circles in the array. These models, however, probably did not provide a fair comparison to human data because of their lack of human-level noise and bias. On the other hand, there exists evidence against heuristic strategies. Im and Halberda (2013) systematically manipulated sample variance ("external variance") and size of the encoding array in a mean size computation task and estimated the number of items needed to be averaged to achieve the observed performance level, assuming a fixed computational error ("internal noise") independent of experimental manipulations. They found that the number of necessary items was seven for their task and suggested the lack of "internal noise" in Myczek and Simons might have led to underestimation of this number. Indeed, a better simulation model should take into account empirically observed psychophysical noise. In the current study, we made use of psychometric functions to estimate the relevant psychophysical noise and imposed it to the simulation of subsampling scenarios. This allowed us to obtain a tighter upper bound of the ideal performance level if subsampling should occur.

Note here we make a distinction between a "heuristic strategy," by which the observers intentionally take advantage of the extremum, range, or a few random samples of the encoding array, and the implicit visual mechanisms by which only a few items can be summarized due to perceptual bottleneck or bottom-up attentional control (Dakin, 2001; Marchant, Simons, & de Fockert, 2013).

Set variance and ensemble perception

Past studies have shown that the visual system is sensitive to not only the mean of a set, but also the variance of it, and these two statistics may be extracted with different levels of efficiency (Solomon, 2010; Solomon, Morgan, & Chubb, 2011). Perception of these two summary statistics may also interact to determine observers' overall perceptual impression of the set (Corbett, Wurnitsch, Schwartz, & Whitney, 2012; Fouriezos, Rubenfeld, & Capstick, 2008; Im & Halberda, 2013). For instance, in Corbett et al. (2012), participants adapted to the mean size of a set of multiple dots and then judged the size of a subsequently presented probing dot. As set variance increased, adaptation effect diminished, suggesting greater difficulty in mean computation. Similarly, Chong and Treisman (2003) asked observers to compare the mean sizes of two simultaneously presented sets of circles and manipulated the size variance of the sets by changing the frequency of each prototypical item in them. Observers were found to perform more accurately when the two sets had the same level of variance than when their variances differed. One possible explanation is that increased heterogeneity would encourage selective attention to individual items in the set. Mean computation has been suggested to be biased toward the subsets that receive more attention (J. W. de Fockert & Marchant, 2008). Alternatively, subsampling occurs during mean computation (Im & Halberda, 2013; Myczek & Simons, 2008). When the set variance is high, the subsampled items would be less representative of the group mean even if these items accounted for a substantial proportion of the entire set.

For facial features, Haberman, Lee, and Whitney (2015) found observers able to accurately estimate the variance in facial expressions of a group, establishing the visual system's sensitivity to variance in high-level domains. We thus asked whether computation of mean facial attractiveness would be affected by set variance. This would not only reveal the interaction of mean computation with other summary statistics, but also characterize the limit of mean computation in terms of robustness against external noise.

The current study

Five experiments were conducted to answer our research questions. In Experiment 1, we constructed and validated a face data set that lies along a continuum of attractiveness. Experiment 2 determined whether participants could compute the mean from a set of heterogeneous faces as well as they could from a set of homogeneous faces. Computational simulations were used to test whether human results could be explained by subsampling one face at random or by sampling the most or the least attractive face systematically from the encoding array. Experiment 3 varied the variance in attractiveness of the encoding array and tested how mean computation would be affected. Experiment 4 was a control experiment to verify that results in Experiments 2 and 3 were due to ensemble perception of the set but not to precise memory of each individual face. Finally, Experiment 5 examined average attractiveness computation using larger arrays of heterogeneous faces, which were asymmetrically distributed around the set mean. Overall, we hypothesized that the visual system could form ensemble perception of attractiveness from a heterogeneous array of faces, and the efficacy of this process would be affected by the group variance in attractiveness. Moreover, we hypothesized that observers' ability to compute the group average would not depend on precise memory of individual members in the group and could be extended to larger and asymmetrical sets.

Experiment 1

In the study of ensemble perception of facial expressions (Haberman & Whitney, 2009), the stimulus faces were from a face data set created by linearly interpolating two faces showing extreme emotions (e.g., happy vs. sad). Our study employed a similar paradigm and similar morphed faces to determine ensemble perception of facial attractiveness. Experiment 1 aimed to construct a data set of facial attractiveness from geometric morphing, which would be used in subsequent experiments to test mean attractiveness perception. To verify that these face stimuli lay on a proper continuum of attractiveness, we presented two faces from the data set simultaneously and tested if the participants' performance in identifying the more attractive face would be a function of the nominal attractiveness contrast between the two faces.

Methods

Participants

Five Chinese undergraduate students (five females; mean age = 22.20, *SD* = 1.92) from the Sun Yat-sen University participated with informed consent and received monetary compensation. All participants had normal or corrected-to-normal vision.

Apparatus and stimuli

Two Chinese female faces with extreme attractiveness from the faces used in Wang, Yao, and Zhou (2015) were hand-segmented to remove the necks, hair, and ears. Average attractiveness was 3.43 and 5.23 for the two faces, respectively, on a nine-point Likert scale from very unattractive to very attractive, rated by 30 male participants (ages: 18–24 years). Geometric morphing with linear interpolation (Sqirlz Morph 2.1, Xibepix, UK) was applied to generate a set of 50 faces from the two extreme faces (Figure 1). We labeled the less attractive one of the two original faces as 1° and the

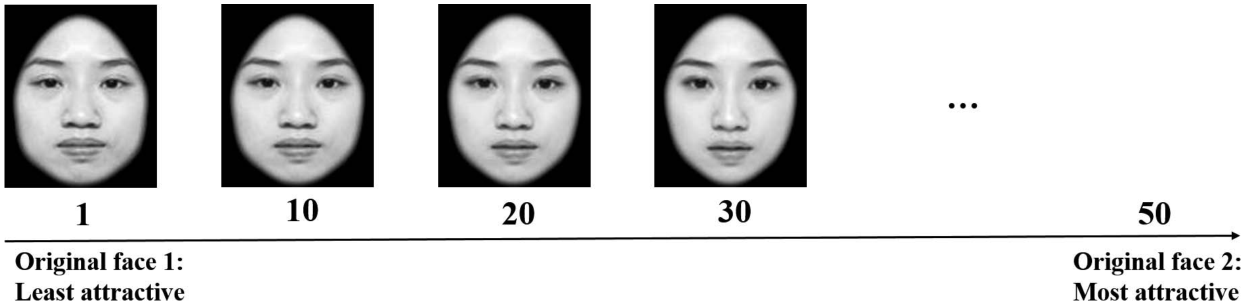

Figure 1. Stimuli used in the current study. Two original faces with extreme attractiveness from the Wang et al. (2015) data set were linearly interpolated to create a set of 50 faces. The least attractive face was labeled as 1° and the most attractive face as 50°. Histogram equalization was applied afterward.

more attractive one as 50° and defined the nominal difference between two consecutive faces in the 50-face data set to be 1°. The 50 faces were then equalized in luminance histogram using the SHINE toolbox (Willenbockel et al., 2010). The final images measured 3.20° × 3.84° (visual degrees) at a viewing distance of 50 cm. All images were presented against a dark background on the 17-in. screen of a Dell computer with a 1,440 × 900 resolution. The experiments were run on MATLAB R2012a with Psychophysics Toolbox 3.0 (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997), and model fitting and data analysis were conducted with MATLAB with Palamedes 1.8.0 (Prins & Kingdom, 2009).

Procedure

In each trial, the participants first fixated a central cross for 500 ms; then they saw two faces from the 50-face data set on the left and right sides of the fixation cross, spatially separated by 1.45° (visual degrees). The two faces differed by 1° to 23° in attractiveness in a step of 2°, and we defined the difference as the *contrast* level (left – right). Whether the face on the left or on the right was more attractive was randomized and balanced. After 2,000 ms, the two faces disappeared, and the participants were instructed to indicate which face was more attractive with a key press. There was no intertrial interval (ITI). After 20 practice trials, participants completed 360 trials (30 trials for each contrast level), which took approximately 15 min. Their response and response time were collected.

Analysis

For each participant, we discarded the trials with a response time greater than 3 *SD* from the individual means (mean discard rate: 1.83% ± 0.73%). For the remaining trials, we define a "positive response" as judging the face on the left as more attractive and a "negative response" otherwise. We used the logistic psychometric function to describe the positive response rate as a function of the contrast level, defined by Wichmann and Hill (2001):

$$\psi(x;\alpha,\beta,\gamma,\lambda) = \gamma + (1-\gamma-\lambda)F(x;\alpha,\beta) \quad (1)$$

with $F(x;\ \alpha,\ \beta)$ being the logistic function:

$$F(x;\alpha,\beta) = \frac{1}{1+exp(-\beta(x-\alpha))}, \quad (2)$$

where x is the contrast level, γ is the guess rate (the probability of giving a positive response when the stimulus is not detected), and λ is the lapse rate (the probability of missing a positive response independent of stimulus intensity). In a two-alternative, forced-choice (2AFC) paradigm, the psychometric function models the proportion of one particular response over the other, and the guess rate γ should be set to zero in model fitting to estimate α, β, and λ (Wichmann & Hill, 2001). In the logistic function, α is the contrast level that corresponds to a 50% positive response rate, i.e., $F(x=\alpha;\ \alpha,\ \beta)=0.5$, which will be used as the threshold of the psychometric function. β is the slope of the psychometric function that indicates the rate of change.

The three free parameters (α, β, and λ) were estimated for each participant using maximum likelihood estimation (MLE). To search for the globally optimal estimates, MLE was run with a grid of different initializing values. We used the log-likelihood ratio to measure the goodness of fit, defined by Wichmann and Hill (2001, equation 5):

$$D = -2log\left(\frac{\mathcal{L}_1}{\mathcal{L}_2}\right), \quad (3)$$

where $\mathcal{L}_1$ is the likelihood of the best-fitting model, and $\mathcal{L}_2$ is the likelihood of the saturated model with no residual errors. The best-fitting model is one that has a unique combination of estimates for α, β, and γ, such that it generates the greatest log-likelihood given the inputs (the contrast level in each trial) and the outputs (the observed positive response rates). The likelihood ($\mathcal{L}_1$ or $\mathcal{L}_2$) is calculated as

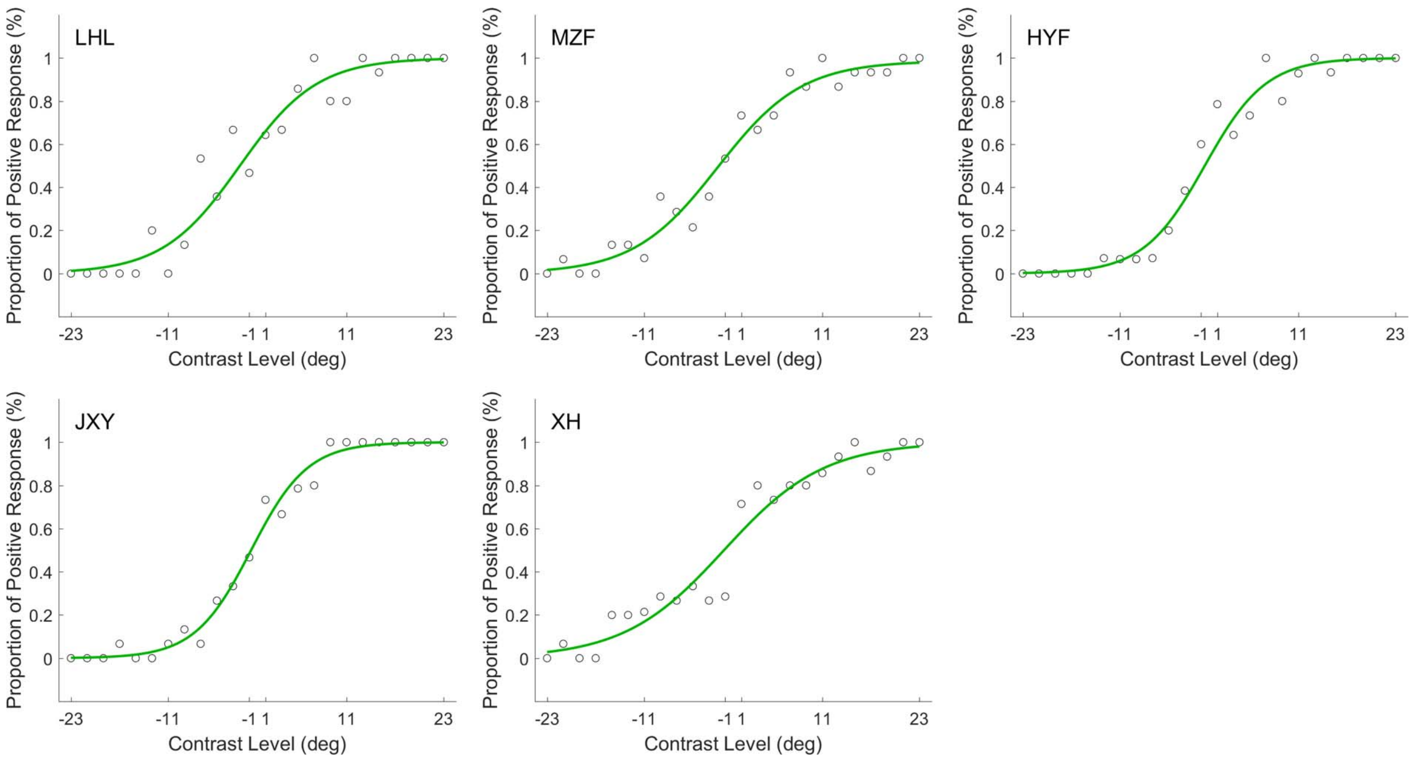

Figure 2. Psychometric functions for each participant in Experiment 1. The proportion of positive response (percentage) is plotted against the attractiveness contrast (degree; left − right). Open circles show the actual response data, and curves show the psychometric functions fitted to the actual data.

$$\mathcal{L} = \prod_{k=1}^{N} p(y_k; x_k, \alpha_0, \beta_0, \gamma_0), \quad (4)$$

where N is the total number of trials, and p is the probability of observing positive response rate y_k given contrast level x_k and the parameter estimates α_0, β_0, and γ_0 in the kth trial. The p value associated with D was estimated by 400 Monte Carlo simulations from the best-fitting model to obtain the distribution of the simulated Ds. The standard errors of the parameters were estimated with 400 runs of bootstrapping.

Finally, the threshold of the psychometric function (the point of subjective equality in our 2AFC paradigm) and the just noticeable difference (JND, the smallest contrast level that can be detected 50% of the time) were used to characterize the overall discriminability level of our stimulus data set. The JND was calculated with the following equation:

$$JND = \frac{t_{75} - t_{25}}{2}, \quad (5)$$

where t_{75} and t_{25} are the contrast levels corresponding to the 75% and 25% positive response rates, respectively, derived from the psychometric function.

Results and discussion

The best-fitting psychometric function for each participant is shown in Figure 2. The estimates of threshold, slope, and lapse rate (see Table 1) were

Participants	Accuracy	Goodness of fit *D* (*p*)	Threshold (*SE*)	Slope (*SE*)	Lapse rate (*SE*)	JND
LHL	84.83%	32.20 (0.045)	−2.16° (0.81°)	0.21° (0.02°)	0.00° (0.01°)	2.27°
MZF	85.27%	17.18 (0.818)	−1.92° (0.96°)	0.19° (0.03°)	0.01° (0.02°)	2.48°
HYF	89.14%	16.47 (0.578)	−0.55° (0.73°)	0.26° (0.03°)	0.00° (0.01°)	1.80°
JXY	89.89%	12.24 (0.765)	−0.77° (0.69°)	0.29° (0.04°)	0.00° (0.01°)	1.66°
XH	84.66%	19.91 (0.698)	−1.15° (1.00°)	0.16° (0.02°)	0.00° (0.02°)	2.95°

Table 1. Performance measures and parameters of psychometric functions in Experiment 1. *Notes*: Goodness of fit was measured by the log-likelihood ratio (D) between the fitted model and the saturated model with no residual errors. The JND was calculated from the fitted psychometric functions. Overall, each participant's performance level was highly comparable, and two faces were discriminable at least 50% of the time when their difference was greater than 2.95°.

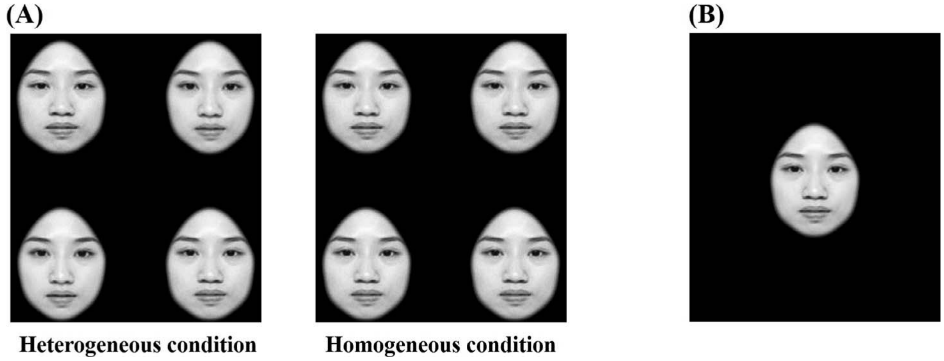

Figure 3. Examples of an encoding array and a probe in Experiment 2. The encoding array (A) consisted of four faces, either identical (the homogeneous condition) or all different, separated from the set mean by ±5° and ±10° (the heterogeneous condition). The probe (B) was a single face.

highly consistent across participants. The mean was −1.31° for the threshold, 0.22° for the slope, and 0.003 for the lapse rate. Overall, goodness-of-fit tests indicated the psychometric function with a small lapse rate and a zero guess rate provided a good fit to the data for all participants (Ds ≤ 19.91, ps ≥ 0.578) except Participant LHL ($D = 32.20$, $p = 0.045$), possibly due to a high level of noise in the data.

The JND (see Table 1) was also comparable across participants with a maximum at 2.95° (Participant XH) and a mean of 2.06° ± 0.39°. Thus, when two simultaneously presented faces were separate by at least 2.95°, all participants could detect their difference at least 50% of the time. Experiment 1 demonstrated that the 50 stimuli produced by geometric morphing lay on a proper attractiveness continuum from which small differences could be sensitively detected.

Experiment 2

In this experiment, we investigated whether participants could compute the mean attractiveness from a set of multiple faces, using the stimulus data set validated in Experiment 1. Specifically, we presented four identical faces (the homogeneous condition) or four different faces (the heterogeneous condition) to participants and asked them to compare the average attractiveness of these four faces with that of a subsequent probe. In light of previous findings of rapid mean computation for facial emotionality (Haberman & Whitney, 2009), gender (Haberman & Whitney, 2007), and identity (J. de Fockert & Wolfenstein, 2009), we hypothesized that participants could determine the average set attractiveness in the homogeneous and the heterogeneous conditions with equivalent sensitivity. Meanwhile, to assess whether participants subsampled one random face or the extrema in the set to perform the task, we used the psychometric function fitted to data in the homogeneous condition to simulate how the participants could have performed in the heterogeneous condition if these subsampling strategies had been used.

Methods

Participants, apparatus, and stimuli

The same five participants as in Experiment 1 and 12 naïve Chinese participants (six females; age: 18–22 years) were recruited for Experiment 2. One of the 12 new participants (female, 19 years old) was found to have misunderstood the instructions and was dropped. All remaining 16 participants had normal or corrected-to-normal vision and gave informed consent. Monetary compensation was given.

The stimuli were identical to those in Experiment 1 except that the encoding array now included four faces placed in a 2 × 2 grid centered on the screen (Figure 3A). The grid occupied an area of 8.17° × 9.50° (visual degrees) viewed at 50 cm. In this experiment and later experiments, all images were presented against a dark background on the 23-in. screen of a Dell computer with a 1,920 × 1,080 resolution. We tested two *encoding conditions*: the *homogeneous* condition, where the four faces were identical, and the *heterogeneous* condition, where they differed from their mean by ±5° and ±10°. Positions of the four faces in the grid were randomly assigned. The probe was a single face placed at the center of the screen (Figure 3B), which differed from the mean of the encoding array by positive or negative 3°, 4°, 8°, 9°, or 12°. This difference defined the *contrast level* (probe – mean). To ensure each contrast level could be tested multiple times against a different set mean, the choice of set means was restricted to between 11° and 40°.

Procedure

On each trial, the participants first fixated a central cross for 500 ms, followed by an encoding array for

2,000 ms. After a blank screen for 500 ms, a probe face was presented until a response was given. Participants were instructed to compute the mean attractiveness of the four faces and indicate whether the probe face was more attractive than the mean or not with a key press. No ITI was given. Here and in other experiments, the participants were not informed about the presence of several different conditions or their nature.

Each participant completed 30 trials per contrast level with a homogeneous condition (one trial per set mean) and 60 trials per contrast level with a heterogeneous condition (two trials per set mean), totaling 450 trials. The encoding conditions were randomized, and whether the probe was more attractive than the set mean was randomly decided on each trial.

Analysis

Trials with a response time greater than 3 *SD* from the individual mean were discarded (mean discard rate: 1.79% ± 0.53%). We defined a "positive response" as judging the probe as more attractive than the set mean. The same psychometric function specified by Equations 1 and 2 were fitted to the positive response rate for each participant and each encoding condition with the independent variable x being the contrast level. Only the threshold (α) and slope (β) parameters indicate properties of the underlying sensory mechanisms. The guess rate (γ) is set to zero, determined by the 2AFC procedure. The lapse rate (λ), which is independent of the stimulus level, can be assumed constant between the two encoding conditions. We thus fixed λ at 0.02 in fitting the psychometric functions. We compared the threshold (α) and slope (β) between the two encoding conditions to examine difference in performance.

Difference in the psychometric functions fitted to the two encoding conditions was examined with a nested hypothesis testing procedure (Mood, Graybill, & Boes, 1974), where the threshold α and the slope β were estimated in three ways: (a) the two encoding conditions fitted with a single function (i.e., one α and one β), (b) the two encoding conditions fitted with β fixed and α free to vary (i.e., two αs and one β), and (c) the two encoding conditions fitted with both α and β free to vary (i.e., two αs and two βs). We computed the log-likelihood ratio $\mathcal{L}_{a,b}$ between models a and b to assess difference in the threshold and the log-likelihood ratio $\mathcal{L}_{b,c}$ between models b and c to assess difference in the slope. Four hundred runs of Monte Carlo simulations were used to estimate distribution of the log-likelihood ratios and the associated p values.

Simulations

We tested the hypotheses of sampling one, two, three, or four faces from the encoding array with computational simulations. The psychometric function in the homogeneous condition describes how the difference between the probe and what is encoded from the four-face array maps to a positive response rate. The same mapping function can be safely assumed for the heterogeneous condition. Based on this assumption, our simulation procedure included four steps: (a) For each participant, we subsampled one to four faces from each actual encoding array the participant had encountered. (b) We then computed the difference between the probe and the sample mean (probe − mean). (c) Using this simulated contrast level, a positive response rate was obtained from the participant's actual psychometric function in the homogeneous condition. These three steps were repeated 10 times to obtain an average positive response rate (p). (d) A Bernoulli sampler then generated a binary response ($k = 0, 1$) with the probability

$$\begin{cases} p, & \text{for } k = 1 \\ q = 1 - p, & \text{for } k = 0 \end{cases}. \quad (6)$$

In this procedure, steps a and b were noise-free. Human-level perceptual noise was added at step c, where the homogeneous psychometric function presumably incorporated both the noise in encoding each individual face and the noise in integrating the faces to compute an average. However, it should be noted that our simulation was likely to provide only an upper bound of the performance level given a subsampling strategy because a certain amount of the noise was not accounted for (e.g., the integration noise might be greater in the heterogeneous condition than in the homogeneous condition, but our simulation assumed they were equal).

Using a similar procedure, we also simulated the mean positive response rate when the attractiveness extremum (the most or the least attractive face in the encoding array) was subsampled.

For all simulations, we tested how much the simulated data deviated from the observed performance level by calculating the negative log-likelihood ratio D, defined in Equation 3, between the likelihood of the actual psychometric function and the likelihood of the saturated model for the simulated data.

Results and discussion

Overall, the 16 participants had a mean accuracy rate of 73.86% (±4.36%) with a mean reaction time of 1.10 s (±0.26 s) per trial (Supplementary Table S1 in Supplementary Materials). The accuracy rates did not differ between the homogeneous and heterogeneous conditions (paired-sample t test), $t(15) = 0.69$, $p = 0.497$, 95% CI = [−0.01, 0.02] for the difference between the means [heterogeneous − homogeneous], Cohen's $d = 0.17$.

The psychometric function described by Equations 1 and 2 provided a good fit to the data, indicated by a small simulated deviation for each participant (Ds $\leq$ 19.70, ps $\geq$ 0.30). Results from a representative participant (Participant HYF) are shown in Figure 4A. Supplementary Table S1 summarizes model-fitting results for each individual participant. Nested hypothesis testing revealed no significant difference in the threshold for 14 out of 16 participants ($\mathcal{L}s_{a,b} \in [0.003, 2.720]$, $ps \in [0.105, 0.957]$), and no significant difference in the slope for 14 out of 16 participants ($\mathcal{L}s_{a,b} \in [0.004, 0.530]$, $ps \in [0.440, 0.957]$). Participants 2-1 and 2-8 had (marginal) differences in the threshold ($\mathcal{L}_{a,b} = 6.57$, $p = 0.014$ for Participant 2-1; $\mathcal{L}_{a,b} = 2.78$, $p = 0.079$ for Participant 2-8), and Participants 2-5 and 2-11 had (marginal) differences in the slope ($\mathcal{L}_{b,c} = 3.60$, $p = 0.067$ for Participant 2-5; $\mathcal{L}_{b,c} = 4.83$, $p = 0.038$ for Participant 2-11). However, combining the effect of threshold and slope, none of these participants showed a significant difference between the two encoding conditions ($\mathcal{L}s \in [2.64, 5.50]$, $ps \in [0.103, 0.323]$). Confirming the subject-based comparisons, paired-sample t tests for all participants combined showed no difference between the homogeneous and heterogeneous conditions for both the threshold, $t(15) = 1.26$, $p = 0.228$, 95% CI = [–2.41, 0.62], Cohen's $d = 0.31$, and the slope, $t(15) = –0.83$, $p = 0.421$, 95% CI = [–0.01, 0.02], Cohen's $d = 0.21$. This result strongly supported our hypothesis that the participants could perform the task in the heterogenous condition as well as they could in the homogeneous condition.

Control analysis

To verify that the participants did not base their response on their knowledge about the probe's position on the attractiveness continuum regardless of the encoding array, we tested whether the positive response rates were affected by the sign of contrast (i.e., "probe < mean" or "probe > mean"). To do so, we took all trials in which the probe was below the midpoint of the attractiveness continuum (i.e., ≤25°) and compared the positive response rates between a subset of these trials in which the probe was below the set mean and in which it was above the set mean. Because some probe degrees had very unbalanced numbers of trials between the "probe > mean" and "probe < mean" conditions, we only used probe degrees that had equal or close-to-equal (differing by one or two) numbers of trials in these two conditions. The resulting probe degrees ranged from 20° to 25°. A paired-sample t test showed a significant difference in positive response rate between "probe < mean" and "probe > mean," $t(15) = 5.27$, $p < 0.001$, 95% CI = [0.13, 0.31], Cohen's $d = 1.32$. Importantly, among these trials, although the probe itself was always below the midpoint of the attractiveness continuum, the positive response rate was still greater when the probe was above the set mean compared to when it was below the set mean (mean difference = 22.26%). Similarly, we performed a t test on probes that were above the midpoint of continuum (ranged between 26° and 31°) and found a significant difference between the "probe > mean" and "probe < mean" conditions, $t(15) = 5.13$, $p < 0.001$, 95% CI = [0.09, 0.22], Cohen's $d = 1.28$. Among these trials, even though the probe was always above the midpoint, the positive response rate was lower when it was below the set mean (mean difference = 15.89%). These results suggested that participants did derive their response from comparison between the probe and the encoding array.

Simulations: Ruling out subsampling strategies

Supplementary Table S2 summarizes how the simulated data compared to the actual performance. Figure 4B through F shows the simulated and observed data from a representative participant (Participant 1). Notably, for every participant, randomly subsampling one face from the encoding array yielded much larger deviance from the observed data than sampling two, three, or four faces. Subsampling the most or the least attractive face among the four resulted in even larger deviance than subsampling one face at random (Supplementary Table S2). In particular, subsampling the most attractive face led to a lower positive response rate at each contrast level, and subsampling the least attractive face led to a higher positive response rate compared to the behavioral data. Overall, simulation results provided evidence against subsampling one face from the array, either at random or choosing the extrema. Consistent with our simulations, the ideal observer model in Myczek and Simons (2008) also performed better when more than one item was sampled.

On the other hand, sampling two, three, and four faces led to lower deviance from the actual data, but their difference seemed to be minimal. Given the amount of noise in the actual data, we were unable to precisely determine which scheme best resembled the actual cognitive mechanism or whether the visual system implicitly alternated between multiple sampling schemes throughout the experiments, depending on the attention and working memory resources available.

Experiment 3

Reduced efficiency or accuracy of ensemble perception in the presence of greater stimulus variance in the

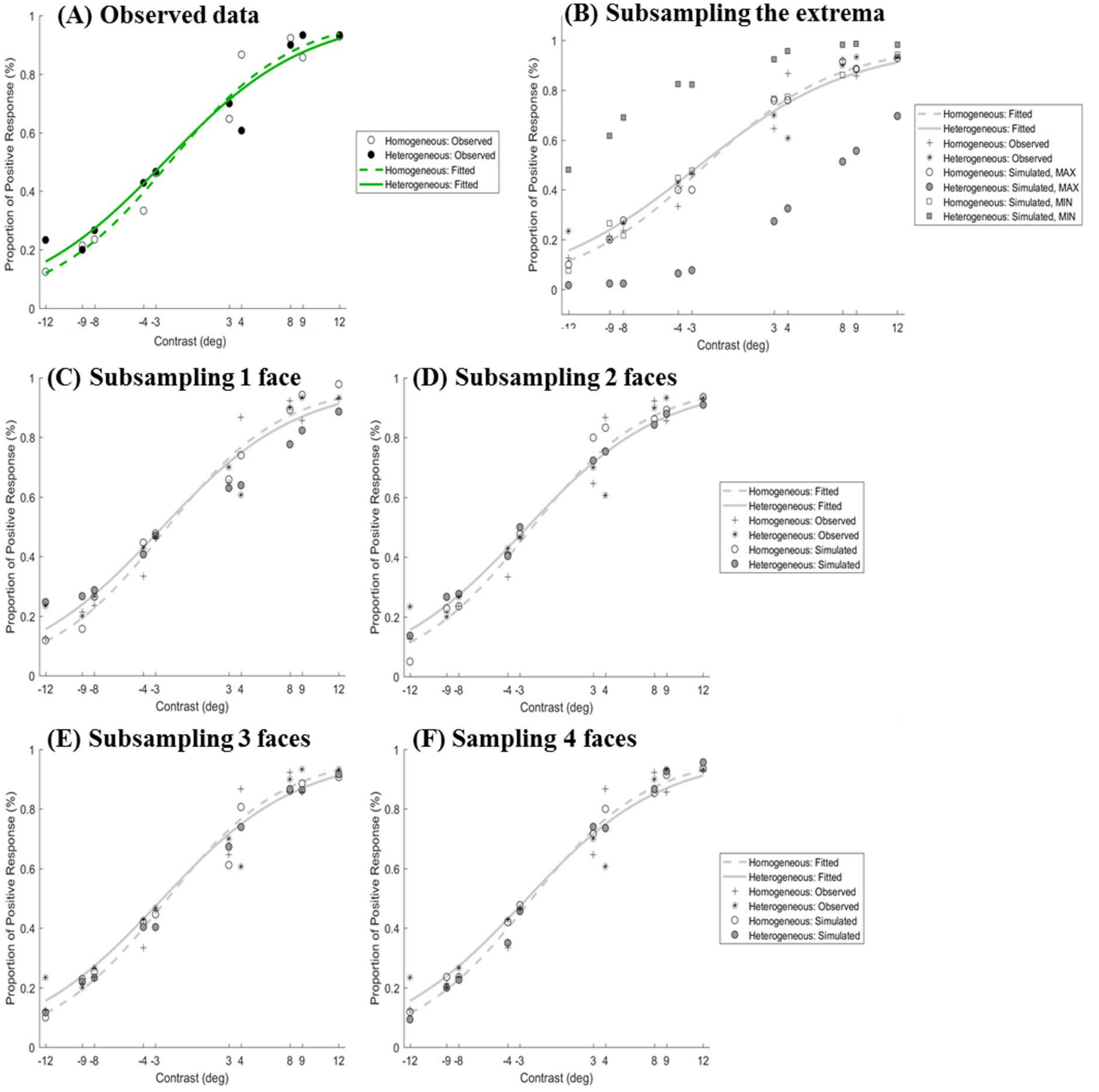

Figure 4. Representative observed and simulated results (Participant HYF) in Experiment 2. (A) The observed response data (open circles = homogeneous condition; filled circles = heterogeneous condition) and the fitted psychometric functions (dashed curve = homogeneous condition; solid curve = heterogeneous condition). Psychometric functions of the homogeneous and heterogeneous conditions showed minimal difference. (B–F) Simulated results are plotted on top of the observed data (+ = homogeneous observed data; * = heterogeneous observed data) and psychometric functions (dashed curve = homogeneous condition; solid curve = heterogeneous condition). (B) The simulated results when the most attractive faces are subsampled (open circles = homogeneous condition; filled circles = heterogeneous condition) or when the least attractive faces (open squares = homogeneous condition; filled squares = heterogeneous condition) are subsampled. (C–F) Simulated results of subsampling one, two, three, or four faces respectively (open circles = homogeneous condition; filled circles = heterogeneous condition).

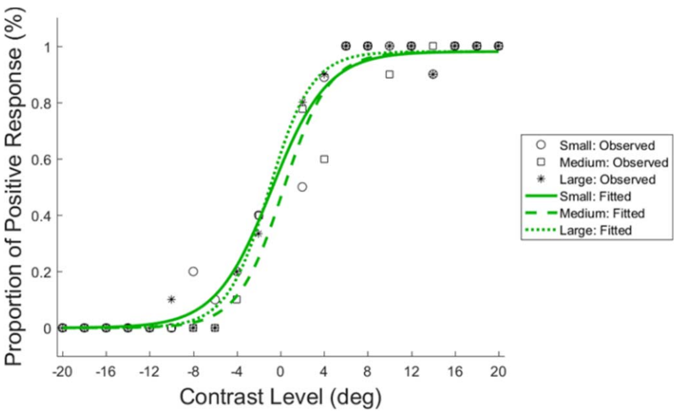

Figure 5. Results for one representative participant (Participant JXY) in Experiment 3. Open circles, open squares, and asterisks show observed percentages of positive response in the small, medium, and large variance conditions, respectively. Solid, dashed, and dotted curves show the fitted psychometric functions in these three variance conditions.

set has been reported for low-level features, such as circle size (Corbett et al., 2012; Im & Halberda, 2013) and orientation (Fouriezos et al., 2008). Would ensemble perception of facial attractiveness be subject to similar effects? Or would it be more resistant to increased local contrast than low-level features? Experiment 3 investigated participants' performance in the ensemble perception task with different levels of set variance. Our hypothesis, following previous findings about ensemble perception of low-level features, was that greater variance in the set would make mean attractiveness computation more difficult.

Methods

Seventeen participants who had participated in Experiment 2, including the one previously excluded, completed Experiment 3 for monetary compensation. The apparatus and stimuli were identical to those in Experiment 2 except that the encoding arrays consisted of four different faces, separated from the mean by ±5° and ±10° (*small* variance condition), ±5° and ±15° (*medium* variance condition), or ±5° and ±20° (*large* variance condition). To ensure the same set means could be tested in all three conditions, we restricted the choice of set mean to the range 21°–30°. The probe was a single face placed at the center of the screen, separated from the mean by positive or negative 2° to 20° in a step of 2°.

On each trial, participants fixated a central cross for 500 ms. An encoding array was shown for 2,000 ms, followed by a blank screen for 500 ms. A probe face then appeared and remained on the screen until a response was given. The participants were instructed to judge whether the probe was more attractive or less attractive than the mean of the four faces in the encoding array. They were given unlimited time to respond with a key press. There was no ITI. Each participant completed 200 trials in each variance condition (10 trials per probe – mean contrast), thus 600 trials in total. All experimental conditions (variance condition, probe condition, and contrast level) were presented in randomized order.

We performed psychometric function fitting and analysis similar to those in Experiment 2 after discarding trials with a response time greater than 3 *SD* from the individual mean (mean discard rate: 2.02% ± 0.68%). Difference in the threshold and the slope between the three variance conditions was examined using nested hypothesis testing.

Results and discussion

The mean accuracy rate for all variance conditions combined was 85.35% ± 4.68% with a mean response time of 0.97 ± 0.20 s per trial. The accuracy rates among the three variance conditions did not differ (repeated-measures ANOVA), $F(2, 32) = 1.88$, $p = 0.169$, $\eta_p^2 = 0.11$.

The psychometric function described by Equations 1 and 2 provided a reasonably good fit to the data (mean deviance Ds ≤ 24.90, ps ≥ 0.168) except for Participant XH in the small variance condition ($D = 24.93$, $p = 0.077$) and Participant 2-8 in the large variance condition ($D = 25.81$, $p = 0.025$) (Supplementary Table S3). The psychometric functions for a representative participant (Participant JXY) are shown in Figure 5. All 17 participants had a nonsignificant difference in the slope across variance conditions ($\mathcal{L}s_{b,c}\in[0.06, 4.04]$, ps$\in[0.151, 0.972]$). In terms of the threshold, 15 out of 17 participants showed nonsignificant difference ($\mathcal{L}s_{a,b}\in[0.07, 4.08]$, ps$\in[0.122, 0.956]$). Participant 2-1 had a significant difference ($\mathcal{L}_{a,b} = 8.66$, $p = 0.012$), and Participant MZF had marginally significant difference ($\mathcal{L}_{a,b} = 4.68$, $p = 0.085$). However, both showed no difference across the three variance conditions if effects of the threshold and the slope were combined in testing ($\mathcal{L} = 6.59$, $p = 0.401$ for Participant 2-1; $\mathcal{L} = 8.75$, $p = 0.414$ for Participant MZF). At the group level, repeated-measures ANOVA showed variance had a nonsignificant effect on the threshold, $F(2, 32) = 1.98$, $p = 0.155$, $\eta_p^2 = 0.11$, but a significant effect on the slope, $F(2, 32) = 4.48$, $p = 0.019$, $\eta_p^2 = 0.22$, which was mainly driven by the difference between the small and the large variance conditions (pairwise t tests with Bonferroni correction: small vs. medium: $p = 1.000$; medium vs. large: $p = 0.149$; small vs. large: $p = 0.077$).

Overall, our data suggested that stimulus variance in the set did not affect ensemble perception of facial attractiveness in a reliable way. Could this result be due to ineffective manipulation of variance? Results in Experiment 1 show that participants' ability to discriminate the attractiveness of any two simultaneously presented faces was a function of the difference between the two faces. Thus, in Experiment 3, the difference in any pair of faces (except the middle pair fixed at $-5°$ and $+5°$ about the mean) would be more noticeable in the large variance condition compared to the medium or the small variance condition. This differential discriminability provided a basis for effective manipulation of variance. Therefore, we conclude that similar performance across variance conditions reflected homogeneity in cognitive processing rather than ineffective manipulation of variance.

Experiment 4

In Experiments 2 and 3, was the participants' high performance level aided by precise memory of individual faces in the encoding array? This would be a crucial confounding factor to rule out because participants might have compared the probe with the remembered faces one by one to derive an accurate response in the ensemble task. Our set size (four faces) fell within the three- to four-item capacity limit of visual working memory for many low-level features, including color, orientation, and bar length (Luck & Vogel, 1997; Zhang & Luck, 2008). More importantly, previous studies suggest working memory capacity might be greater than three or four items for facial identities (Jiang, Shim, & Makovski, 2008) and could be further enhanced if the faces were familiar to the observers (Jackson & Raymond, 2008). Because our face data set was comprised of two unique identities and their morphs, the likelihood of the participants precisely remembering the four encoding faces on each trial could be high. To verify this was not the case, we directly examined participants' memory of the individual faces in Experiment 4.

On the other hand, the ability to form ensemble perception of a set without holding precise information about its individual members has been consistently observed in many feature domains, including circle size (Ariely, 2001), orientation (Parkes et al., 2001), and facial expressions (Haberman & Whitney, 2009). We thus hypothesized that, at the same stimulus set size, participants' memory of individual faces would not be able to account for their high performance level in the ensemble task.

Methods

Participants, apparatus, and stimuli were the same as in Experiment 3. Participants first fixated a central cross for 500 ms before they viewed an encoding array for 2,000 ms. The encoding arrays were identical to those in Experiment 3, but the participants were now instructed to remember the four faces rather than to compute their mean attractiveness. After a blank screen for 500 ms, the probe appeared, which consisted of a single face that was either one of the four encoding faces (*member* condition) or a new face that differed from any face in the encoding set by at least 7° (*nonmember* condition). Participants indicated with a key press whether the probe was a member of the encoding array or not. Response time was unlimited, and the probe remained on the screen until a response was given. There was no ITI.

Same as in Experiment 3, three variance conditions were tested. In each variance condition, the participants completed 40 trials of the *member* condition and 80 trials of the *nonmember* condition. Each participant completed 360 trials in total.

Trials with a response time greater than 3 *SD* from the individual mean were discarded (mean discard rate: 2.07% ± 0.93%). We used d' as a metric to compare participants' sensitivity in the ensemble perception task (Experiment 3) and in the working memory task (the current experiment). In the working memory task, we define the *hit* rate as the proportion of the *member* trials being correctly identified and the *false alarm* rate as the proportion of the *nonmember* trials being mistaken as a *member*. The d' is computed as

$$d' = \Phi^{-1}(p_{HIT}) - \Phi^{-1}(p_{FA}), \quad (7)$$

where Φ^{-1} is the normal inverse cumulative distribution, p_{HIT} is the hit rate, and p_{FA} is the false alarm rate.

For the ensemble task, we reused data from Experiment 3 and defined the *hit* rate for this task as the proportion of trials in which the probe face is more attractive than the average and the participant correctly identifies the probe as more attractive and the *false alarm* rate as the proportion of trials in which the probe is less attractive than the average, but the participant mistakenly reports it as more attractive. d' is computed following Equation 7.

Results and discussion

Overall, the mean accuracy rate in the working memory task was 54.19% ± 3.54%, and the mean response time was 1.09 ± 0.39 s per trial. The accuracy rates differed significantly among variance conditions, $F(2, 32) = 21.14$, $p < 0.001$, $\eta_p^2 = 0.57$.

A 2 (task conditions: ensemble perception vs. working memory recall) × 3 (variance conditions: small vs. medium vs. large) repeated-measures ANOVA on d' showed significant main effects of task, $F(1, 16) = 487.75$, $p < 0.001$, $\eta_p^2 = 0.88$, and variance, $F(2, 32) =$

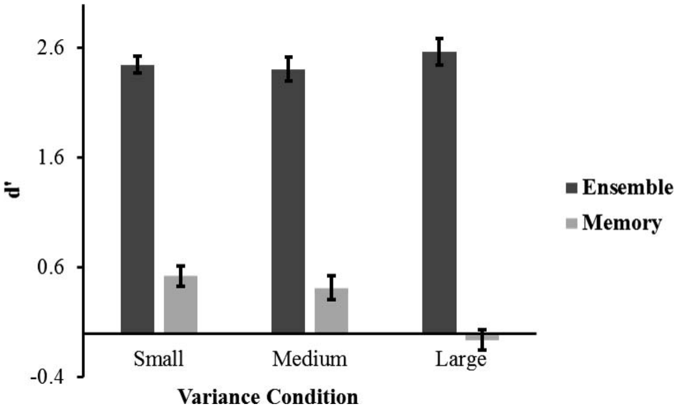

Figure 6. *d*′ in Experiments 3 and 4. The dark gray bars show *d*′ in the ensemble task (Experiment 3), and the light gray bars show *d*′ in the memory task (Experiment 4). The main effect of Task was significant. Error bars indicate ±1 *SEM*.

4.51, $p = 0.002$, $\eta_p^2 = 0.06$ (Figure 6). Further *t* tests revealed *d*′ in the ensemble perception task was greater than that in the working memory task at all variance levels, *t*s(16) ≥ 13.96, *p*s < 0.001, Cohen's *d*s ≥ 3.39.

An interaction between task condition and variance condition was also observed, $F(2,32) = 10.25$, $p < 0.001$, $\eta_p^2 = 0.13$, suggesting that working memory and ensemble perception were affected by the variance condition differently. Consistent with this, a one-way repeated-measures ANOVA showed variance condition did not affect *d*′ in the ensemble task, $F(2, 32) = 1.76$, $p = 0.188$, $\eta_p^2 = 0.03$, but did in the memory task, $F(2, 32) = 9.86$, $p < 0.001$, $\eta_p^2 = 0.30$. For *d*′ in the memory task, post hoc paired *t* tests with Bonferroni correction showed a difference between the small and the large variance conditions, $t(16) = 4.11$, $p = 0.027$, Cohen's $d = 1.00$, 95% $CI = [0.28, 0.88]$, and between the medium and the large variance conditions, $t(16) = 2.97$, $p = 0.003$, Cohen's $d = 0.72$, 95% $CI = [0.14, 0.82]$. In fact, mean *d*′ was not different from chance level zero in the large variance condition, $t(16) = –0.65$, $p = 0.522$, Cohen's $d = 0.16$, 95% $CI = [–0.25, 0.13]$; but was significantly different from zero in the small and medium variance conditions, *t*s(16) ≥ 3.91, *p* ≤ 0.001, Cohen's *d*s ≥ 0.95.

Taken together, sensitivity in the working memory task was lower than that in the ensemble task, and set variance affected working memory but not ensemble perception of the four faces. Thus, the high performance level in the ensemble task could not be fully accounted for by working memory of individual faces in the encoding array. Other perceptual mechanisms specific to ensemble encoding were required.

Interestingly, working memory of sets with large variance was significantly worse than sets with smaller variance. One possibility would be that participants automatically computed the set average, and their perception of individual faces in the set was biased toward this average. In the large variance condition, the difference between the set average and each individual face was larger than in the small and medium variance conditions; thus, the bias led to greater memory error. Indeed, summary statistics can influence perception of embedded individual objects. For instance, in a multicircle array, working memory of the size of an individual circle was found to be biased toward the mean size of all circles in the same color and to the mean size of all circles in the array (Brady & Alvarez, 2011).

Experiment 5

In Experiments 2 and 3, the encoding arrays were always presented for 2,000 ms, fixed at a set size of four faces, and contained faces that were symmetrically distributed about the set mean. Moreover, in Experiment 3, we only manipulated the two faces with extreme attractiveness values in the array and fixed the nonextreme faces at ±5° about the set mean. These could potentially encourage encoding heuristics, such as (a) subsampling the two faces with extreme attractiveness values in the array and (b) remembering the attractiveness values of the individual faces, and relied on nonvisual information for response. To rule out these possibilities and to test whether perception of average attractiveness could be generalized to arrays with a greater set size and with asymmetrical distribution of attractiveness degrees around the mean, we conducted Experiment 5, in which we introduced two further manipulations: First, we increased the size of the encoding arrays from four (one instance per prototype) to eight (two instances per prototype) or 12 (three instances per prototype) while fixing the exposure time at 2,000 ms; second, we manipulated attractiveness of the nonextreme faces in an array while holding the extreme faces constant. Based on previous findings that increasing the set size only marginally affected perception of average circle size (Ariely, 2001) and average facial emotionality (Haberman & Whitney, 2009), we hypothesized that participants' judgment of the average attractiveness would not be affected significantly by the set size. Meanwhile, we hypothesized that manipulating distribution of the faces would skew participants' judgment of the average, suggesting they do not base their judgment only on the extreme faces; faces in the middle range would also need to be sampled.

Methods

Twenty naïve Chinese participants were recruited for Experiment 5 (11 females; age: 18–22 years), who gave informed consent and were compensated with monetary rewards. The apparatus and stimuli were identical to those in Experiment 2. The encoding array included four prototype faces, and we changed the number of instances per prototype to manipulate the set size: set sizes four, eight, and 12 contained one, two, and three instances per prototype, respectively. Faces in the encoding array were placed on a circle (radius = 10.88°) around the central fixation cross, equally spaced. In choosing the four prototype faces, rather than placing the four faces around the set average symmetrically (such as the conditions in Experiments 2 and 3), we introduced two bias conditions: In the *positive bias* condition, the four faces were at –20°, +5°, +15°, and +20° around the nominal set mean, which is the average of the two extreme faces (±20°); in the *negative bias* condition, the four faces were at –20°, –15°, –5°, and +20° around the nominal set mean. Thus, in the positive bias condition, the true set mean was more attractive than the nominal set mean and vice versa in the negative bias condition. The nominal means were chosen in the range 21°–30° on the attractiveness continuum. The probe face was separated from the nominal set mean by positive or negative 3°, 4°, 8°, 9°, or 12°. Note that this made the negative bias condition have more trials with the probe more attractive than the true set mean and vice versa for the positive bias condition. However, the psychometric functions, fitted against the contrast between the probe and the nominal contrast, would not be affected by this asymmetry.

The experimental procedure was the same as in Experiment 3 regardless of set size of the encoding array. All set size × bias conditions were presented in random order. Each participant completed 100 trials in each set size × bias condition, totaling 600 trials. Trials with a response time greater than 3 *SD* from the individual mean were discarded (mean discard rate = 1.67% ± 0.67%). For each set size × bias condition, we fitted a psychometric function to the positive response rate against the contrast between the probe and the nominal set mean (probe – nominal mean) following the procedure described in Experiment 2.

Results and discussion

The mean response accuracy (with respect to the true mean) was 73.62% ± 3.78% with an average response time of 1.12 ± 0.36 s per trial. The accuracy rates were affected by the bias, $F(1, 19) = 52.51$, $p < 0.001$, $\eta_p^2 = 0.73$, with better accuracy in the negative bias condition (mean accuracy = 79.90%) than in the positive bias condition (mean accuracy = 67.37%). This was likely to be caused by unbalanced distribution of the probe around the true set mean between the two bias conditions: Because participants might be biased toward judging the probe as more attractive (as evidenced by a negative threshold for most psychometric functions in Experiments 2 and 3) in the negative bias condition in which more trials had a probe greater than the true set mean, their judgment was more likely to be correct. But the accuracy rates were not affected by the set size, $F(2, 38) = 0.89$, $p = 0.421$, $\eta_p^2 = 0.32$, or its interaction with the bias, $F(2, 38) = 1.48$, $p = 0.240$, $\eta_p^2 = 0.07$.

The psychometric function described by Equations 1 and 2 provided a good fit to the data (deviance ps ≥ 0.638; see Figure 7 for plots from two representative participants). Supplementary Table S4 summarizes psychometric model fitting for each individual participant. A two-way repeated-measures ANOVA on the threshold (Figure 8) revealed a significant main effect of the bias condition, $F(1, 19) = 47.22$, $p < 0.001$, $\eta_P^2 = 0.71$, and marginal main effect of set size, $F(2, 38) = 2.74$, $p = 0.077$, $\eta_P^2 = 0.13$. The interaction effect was nonsignificant, $F(2, 38) = 0.38$, $p = 0.687$, $\eta_P^2 = 0.02$. Regarding the marginal effect of set size, post hoc pairwise t tests (with Bonferroni correction) showed *n.s.* difference in the threshold between set sizes four and eight ($p = 1.000$) or between set sizes eight and 12 ($p = 0.254$). The main difference was between set sizes four and 12 ($p = 0.048$). Importantly, when we examined the effect of bias at each set size independently, paired-sample t tests revealed a significant difference between the two bias conditions at all set sizes, set size four: $t(19) = 5.35$, $p < 0.001$, 95% CI = [1.12, 2.76], Cohen's $d = 1.20$; set size eight: $t(19) = 4.64$, $p < 0.001$, 95% CI = [9.98, 2.69], Cohen's $d = 1.04$; set size 12: $t(19) = 4.68$, $p < 0.001$, 95% CI = [0.88, 2.31], Cohen's $d = 1.05$, when the difference was computed as "positive bias condition – negative bias condition". In contrast, the slope (Figure 8) did not seem to be affected by either the bias conditions, $F(1, 19) = 0.39$, $p = 0.541$, $\eta_P^2 = 0.02$; the set size, $F(2, 38) = 0.64$, $p = 0.532$, $\eta_P^2 = 0.03$; or their interaction, $F(2, 38) = 0.56$, $p = 0.575$, $\eta_P^2 = 0.03$. Bias did not affect the slope at each set size independently (ps ≥ 0.245).

Overall, in this experiment, we observed that participants' performance and sensitivity were significantly affected by the bias condition. A positive bias resulted in larger thresholds than a negative bias (from the 95% confidence intervals in the paired-sample t tests) regardless of the set size. This is consistent with the hypothesis that when the middle faces skew the set mean to a more positive value, participants are less likely to judge the probe as more attractive than the (true) set mean than when the middle faces skew the set mean to a more negative value. Thus, not only the extreme faces, but also faces in the middle of the range, are pooled when average attractiveness is computed.

On the other hand, we observed that a larger set size somehow altered observers' sensitivity to the stimuli

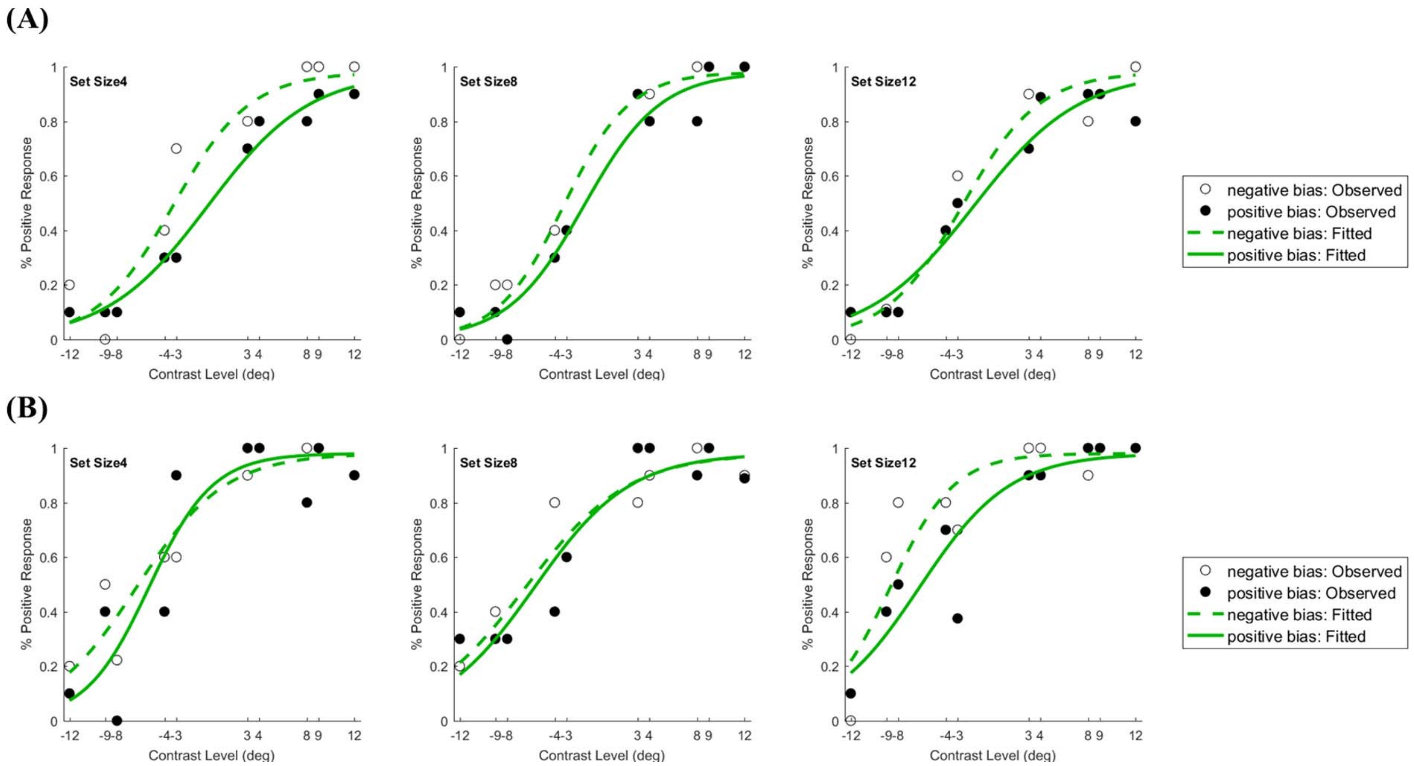

Figure 7. Fitted psychometric functions for (A) Participant 5-3 and (B) Participant 5-16. For each of Panels A and B, the left, middle, and right panels show psychometric curves at set sizes four, eight, and 12, respectively.

(marginal effects on the threshold and the slope), but such a change in sensitivity had little effect on performance accuracy. What caused this change in stimulus sensitivity? One possible explanation could be that faces in a set were pooled serially during computation of average attractiveness, and a larger set size reduced the amount of time allowed to process each face. However, because attractiveness information of single faces could be extracted from 100 ms of presentation time (Locher, Unger, Sociedade, & Wahl, 1993), the 2,000-ms presentation time in our experiment could still allow rapid serial encoding of all individual faces even in the largest sets (12 faces). Alternatively, faces in a set might be pooled in parallel, but mean computation might be performed with different levels of efficiency at different set sizes.

General discussion

In four experiments, we demonstrated a novel phenomenon: Ensemble perception of attractiveness from a group of faces. Experiment 1 verified that our 50-face data set represents a valid attractiveness continuum on which the difference in attractiveness of

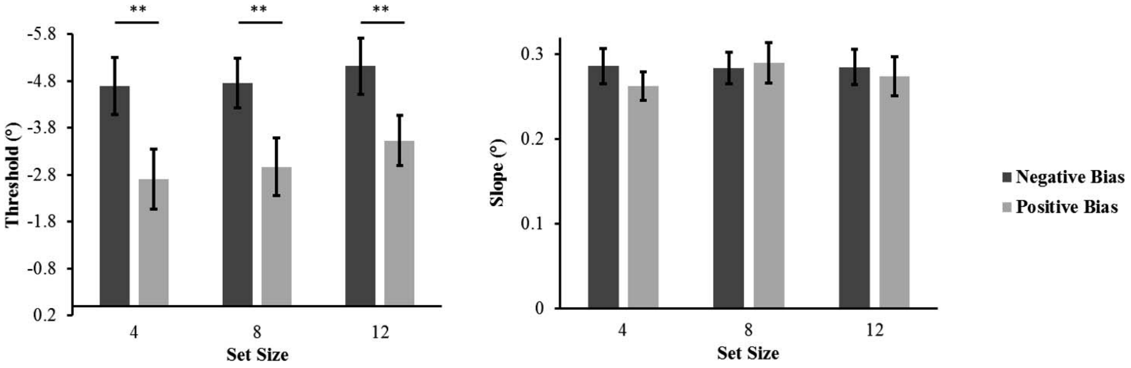

Figure 8. The mean threshold and slope in each bias × set size condition. Overall, the threshold was significantly affected by the bias and marginally affected by the set size. The slope was not affected by either factor. Error bars indicated ±1 *SEM*. $**p < 0.001$.

two juxtaposed faces can be detected 50% of the time when they are at least 2.95° apart. Using this data set, Experiment 2 found that the participants could compute the mean attractiveness from multiple faces showing different degrees of attractiveness as well as they could from multiple faces showing the average attractiveness. Simulation results suggested their performance could not be explained by randomly choosing one face or systematically choosing the most or the least attractive face from the array for comparison with the probe. Experiment 3 manipulated the variance in attractiveness in the array and found mean computation was unaffected. Experiment 4 suggested participants did not rely on memory of individual faces in the encoding array to perform well in the ensemble perception task. Finally, Experiment 5 showed that computation of average attractiveness could be extended to larger set sizes and encoding arrays in which the composing faces were distributed asymmetrically around the set mean.

Ensemble perception of facial attractiveness

The finding of average facial attractiveness perception in our study agrees with and extends previous studies that demonstrate ensemble perception of high-level complex features (J. de Fockert & Wolfenstein, 2009; J. W. de Fockert & Gautrey, 2013; Haberman & Whitney, 2007; Haberman & Whitney, 2009; Neumann et al., 2013).

Our finding that group-level impression of heterogeneous faces is equivalent to that of repeats of the set average might offer some implications for the cheerleader effect (Walker & Vul, 2014), the assimilation effect (Geiselman, Haight, & Kimata, 1984), and the group attractiveness effect (van Osch et al., 2015). To explain the cheerleader effect, Walker and Vul (2014) propose that observers could have mentally morphed multiple faces in a group, and this morphed face is perceived as more attractive than most of its compositing faces due to its averageness (Langlois & Roggman, 1990). Observers' subsequent judgment of an individual face is biased toward the attractiveness of this morphed face. Thus, an individual face in a group is perceived to be more attractive than when the face is presented alone. Our findings partially support this hypothesis. On one hand, we demonstrated the possibility of extracting an ensemble representation of a heterogeneous group of faces. On the other hand, if the ensemble representation is perceived as more attractive than most of the compositing faces (as Walker and Vul speculate), we would expect to see participants more often judge the ensemble representation as more attractive than the arithmetic mean of the compositing faces. However, in Experiments 2 and 3, the threshold estimates were predominantly negative, suggesting that the participants were *not* biased toward judging the ensemble representation to be more attractive than the arithmetic mean. Attractive faces are not always average (DeBruine, Jones, Unger, Little, & Feinberg, 2007; Halberstadt, Pecher, Zeelenberg, Ip Wai, & Winkielman, 2013). This casts doubt on Walker and Vul's premise and suggests an alternative explanation is needed for the cheerleader effect.

Similarly, participants' tendency to summarize the average attractiveness of the displayed face set could also provide a new perspective on the assimilation effect (Geiselman et al., 1984), in which a target face is perceived as more attractive when it is surrounded by high-attractiveness faces and less attractive when it is surrounded by low-attractiveness faces. When a target face is surrounded by a group of faces with heterogeneous attractiveness, observers might implicitly extract the average attractiveness of the group, and their perception of the target face could be biased toward the group attractiveness.

On the other hand, van Osch et al. (2015) suggested the group attractiveness effect could be due to selective attention to the more attractive members in the set. However, we did not find the effect of selective attention on ensemble attractiveness perception with the paradigm in our experiments as simulations in Experiment 2 that subsampled the most attractive face in the encoding array resulted in large deviance from the observed data. Moreover, in the medium and large variance conditions of Experiment 3, there was a higher likelihood of the most attractive face in a set being more attractive than the probe compared to the small variance condition. But this did not differentiate performance level in the three conditions. Nevertheless, as van Osch et al. noted, the group attractiveness effect diminished in smaller groups of four or five people or when participants paid focused attention to each group member. It is thus possible that the perceptual bottleneck of processing multiple faces had contributed to the group attractiveness effect and affect the set variance effect here (see detailed discussion below).

How many faces are sampled?

Simulation results in Experiment 2 demonstrated that observers were able to pool at least two faces from the four-face array in computing the average attractiveness. Moreover, Experiment 5 expanded the set size to eight and 12 faces, which made only marginal, if any, difference in the performance and psychometric parameters compared to the four-face arrays. This suggested it was very likely that more than two faces were integrated when the array contained eight or 12 faces; otherwise, because two-face subsampling would be less representative of the set average as the set size increased, we should have seen a clear change in performance level and psychometric parameters. These results are in line with experimental and modeling studies that showed multiple items need to be integrated for ensemble perception and consistent with the finding that the number of items pooled would be approximately the square root of the set size (Dakin, 2001; Dakin, Bex, Cass, & Watt, 2009). Similar to our results, using ideal-observer models with empirically derived noise in size estimation, Sweeny, Wurnitsch, Gopnik, and Whitney (2015) showed that young children (4–5 years in age) integrated at least four items and adults sampled at least seven items across two simultaneously presented arrays (each with eight items) in a task to compare their average sizes. Sweeny and Whitney (2014) performed similar noisy subsampling stimulations in ensemble perception of a crowd's gaze and found that observers pooled at least two from the four-face array. However, it should also be noted that, although the visual system has the capability of pooling multiple items from the visual field in a parallel fashion for averaging (Im & Halberda, 2013; Solomon, 2010), it is also possible to be influenced by individual items that receive overt or covert selective attention (Chong & Treisman, 2005; J. W. de Fockert & Marchant, 2008).

Set variance

In our Experiment 3, set variance in attractiveness among faces did not significantly affect average attractiveness computation. This result does not agree with a previous finding that increasing set variance impaired ensemble perception (Corbett et al., 2012; Fouriezos et al., 2008; Im & Halberda, 2013; Marchant et al., 2013; Solomon et al., 2011).

We suggest that the discrepancy might be due to the fact that the variance effect on ensemble perception may be more prominent when the perceptual inputs exceed the capacity limit of focal attention. Using circles as stimuli, Marchant et al. (2013) found a smaller variance effect at set size four or eight than at set size 16. In fact, performance difference between small variance and large variance sets at set size four was barely discernible, likely to be nonsignificant (figure 4 of Marchant et al., 2013; no statistical test was run on this). Set size four, which is employed by Experiment 3 of the current study, presumably falls within the limit of focal attention (Pylyshyn & Storm, 1988). By contrast, most studies that observed a variance effect used a larger set size (e.g., 14 items per set in Corbett et al., 2012). When the information to be encoded falls outside focal attention, perception of the mean might become more susceptible to the influence of other summary statistics, such as the variance.

Another possible explanation for the lack of variance effect on averaging facial attractiveness is that our study used a high-level complex feature, and the other studies tested ensemble perception with low-level features, such as object size (Corbett et al., 2012; Im & Halberda, 2013; Marchant et al., 2013; Solomon et al., 2011) or orientation (Fouriezos et al., 2008). The underlying mechanism of ensemble perception of high-level features and that of low-level features might be different as Haberman, Brady, and Alvarez (2015) found that low-level ensemble representations (e.g., orientation and color) correlated well with each other but not with high-level ensemble representations (e.g., facial expression and person identity).

Future directions and limitations

Ensemble perception of facial attractiveness may be of evolutional importance as it enables comparison of attractiveness on a group basis for socially significant inferences and decision making (Phillips, Weisbuch, & Ambady, 2014). Moreover, our impression of the overall attractiveness of a social group could implicitly influence our judgment of individual members in the group, which could subsequently bias inference of personal traits and likeability of individual members. To explore whether and how ensemble perception of facial attractiveness affects group (member) social perception (e.g., stereotype) is a promising future research direction.

As for the limitations, our experiments used only female faces as stimuli. The sexual orientations of our participants in the present study were not recorded although most Chinese students identify as heterosexual. The participants and stimuli shared the same racial background. Racial and gender information of the stimulus faces is found to influence judgment of facial attractiveness (Keating, 1985; Rhodes et al., 2005). It is thus possible that these traits also influence perception of average attractiveness of a group, and our results may not readily extend to other stimulus conditions, such as other-race or mixed-race composite faces.

Conclusion

To conclude, we found observers capable of extracting an ensemble perception of attractiveness from multiple faces. We also took the initial steps to characterize this phenomenon with respect to subsampling and stimulus variance. Our findings expand the feature domains with which ensemble perception can operate and motivate further investigation of the cognitive and neural basis and social implications of this cognitive process.

Keywords: facial attractiveness, ensemble perception, statistic summary, the cheerleader effect, the group attractiveness effect

Acknowledgments

This research was supported by the National Natural Science Foundation of China (31170978, 31771208), the Major Program of the National Social Science Foundation of China (14ZDB015), and the Fundamental Research Funds for the Central Universities (17wkzd19). We thank Pan Zhao, Fei Lin, Lunyan Chen, and Lei Tang for collecting the data.

Commercial relationships: none.
Corresponding author: Guomei Zhou.
Email: zhougm@mail.sysu.edu.cn.
Address: Department of Psychology, Sun Yat-sen University, Guangzhou, China.

References

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

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

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, *12*(2), 157–162.

Bauer, B. (2009). Does Stevens's power law for brightness extend to perceptual brightness averaging? *Psychological Record*, *59*(2), 171–185.

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, https://doi.org/10.1177/0956797610397956.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436.

Cataliotti, J., & Gilchrist, A. (1995). Local and global processes in surface lightness perception. *Perception & Psychophysics*, *57*(2), 125–135.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research*, *43*(4), 393–404.

Chong, S. C., & Treisman, A. (2005). Attentional spread in the statistical processing of visual displays. *Perception & Psychophysics*, *67*(1), 1–13.

Corbett, J. E., Wurnitsch, N., Schwartz, A., & Whitney, D. (2012). An aftereffect of adaptation to mean size. *Visual Cognition*, *20*(2), 211–231, https://doi.org/10.1080/13506285.2012.657261.

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

Dakin, S. C., Bex, P. J., Cass, J. R., & Watt, R. J. (2009). Dissociable effects of attention and crowding on orientation averaging. *Journal of Vision*, *9*(11):28, 1–16, https://doi.org/10.1167/9.11.28. [PubMed] [Article]

DeBruine, L. M., Jones, B. C., Unger, L., Little, A. C., & Feinberg, D. R. (2007). Dissociating averageness and attractiveness: Attractive faces are not always average. *Journal of Experimental Psychology, Human Perception and Performance*, *33*(6), 1420–1430, https://doi.org/10.1037/0096-1523.33.6.1420.

de Fockert, J., & Wolfenstein, C. (2009). Rapid extraction of mean identity from sets of faces. *Quarterly Journal of Experimental Psychology*, *62*(9), 1716–1722, https://doi.org/10.1080/17470210902811249.

de Fockert, J. W., & Gautrey, B. (2013). Greater visual averaging of face identity for own-gender faces. *Psychonomic Bulletin & Review*, *20*(3), 468–473, https://doi.org/10.3758/s13423-013-0381-8.

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

Fouriezos, G., Rubenfeld, S., & Capstick, G. (2008). Visual statistical decisions. *Perception & Psychophysics*, *70*(3), 456–464, https://doi.org/10.3758/PP.70.3.456.

Geiselman, R. E., Haight, N. A., & Kimata, L. G. (1984). Context effects on the perceived physical attractiveness of faces. *Journal of Experimental Social Psychology*, *20*(5), 409–424, https://doi.org/10.1016/0022-1031(84)90035-0.

Geisler, W. S., Perry, J. S., Super, B. J., & Gallogly, D. P. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research*, *41*(6), 711–724.

Grammer, K., & Thornhill, R. (1994). Human (Homo sapiens) facial attractiveness and sexual selection: The role of symmetry and averageness. *Journal of Comparative Psychology*, *108*(3), 233–242.

Griffin, A. M., & Langlois, J. H. (2006). Stereotype directionality and attractiveness stereotyping: Is beauty good or is ugly bad? *Social Cognition*, *24*(2), 187–206, https://doi.org/10.1521/soco.2006.24.2.187.

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., Lee, P., & Whitney, D. (2015). Mixed emotions: Sensitivity to facial variance in a crowd of faces. *Journal of Vision*, *15*(4):16, 1–11, https://doi.org/10.1167/15.4.16. [PubMed] [Article]

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

Haberman, J., & Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(3), 718–734, https://doi.org/10.1037/a0013899.

Halberstadt, J., Pecher, D., Zeelenberg, R., Ip Wai, L., & Winkielman, P. (2013). Two faces of attractiveness: Making beauty in averageness appear and reverse. *Psychological Science*, *24*(11), 2343–2346, https://doi.org/10.1177/0956797613491969.

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, https://doi.org/10.3758/s13414-012-0399-4.

Jackson, M. C., & Raymond, J. E. (2008). Familiarity enhances visual working memory for faces. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(3), 556–568, https://doi.org/10.1037/0096-1523.34.3.556.

Jiang, Y. V., Shim, W. M., & Makovski, T. (2008). Visual working memory for line orientations and face identities. *Perception & Psychophysics*, *70*(8), 1581–1591, https://doi.org/10.3758/PP.70.8.1581.

Keating, C. F. (1985). Gender and the physiognomy of dominance and attractiveness. *Social Psychology Quarterly*, *48*(1), 61–70.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3. *Perception*, *36*(14), 1–16.

Langlois, J. H., & Roggman, L. A. (1990). Attractive faces are only average. *Psychological Science*, *1*(2), 115–121, https://doi.org/10.1111/j.1467-9280.1990.tb00079.x.

Locher, P., Unger, R., Sociedade, P., & Wahl, J. (1993). At first glance: Accessibility of the physical attractiveness stereotype. *Sex Roles*, *28*(11/12), 729–743.

Luck, S. J., & Vogel, E. K. (1997, November 20). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281, https://doi.org/10.1038/36846.

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

Mood, A. M., Graybill, F., & Boes, D. C. (1974). *Introduction to the theory of statistics* (3rd ed.). New York: McGraw-Hill.

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

Neumann, M. F., Schweinberger, S. R., & Burton, A. M. (2013). Viewers extract mean and individual identity from sets of famous faces. *Cognition*, *128*(1), 56–63, https://doi.org/10.1016/j.cognition.2013.03.006.

Nothdurft, H. C. (1991). Texture segmentation and pop-out from orientation contrast. *Vision Research*, *31*(6), 1073–1078.

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

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442.

Phillips, L. T., Weisbuch, M., & Ambady, N. (2014). People perception: Social vision of groups and consequences for organizing and interacting. *Research in Organizational Behavior*, *34*, 101–127, https://doi.org/10.1016/j.riob.2014.10.001.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, *40*(1), 49–70, https://doi.org/10.1023/A:1026553619983.

Prins, N., & Kingdom, F. A. A. (2009). Palamedes: Matlab routines for analyzing psychophysical data. Retrieved from http://www.palamedestoolbox.org/

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision*, *3*(3), 179–197.

Rhodes, G., Lee, K., Palermo, R., Weiss, M., Yoshikawa, S., Clissa, P., . . . Jeffery, L. (2005). Attractiveness of own-race, other-race, and mixed-race faces. *Perception*, *34*(3), 319–340.

Solomon, J. A. (2010). Visual discrimination of orientation statistics in crowded and uncrowded arrays. *Journal of Vision*, *10*(14):19, 1–16, https://doi.org/10.1167/10.14.19. [PubMed] [Article]

Solomon, J. A., Morgan, M., & Chubb, C. (2011). Efficiencies for the statistics of size discrimination. *Journal of Vision*, *11*(12):13, 1–11, https://doi.org/10.1167/11.12.13. [PubMed] [Article]

Sweeny, T. D., & Whitney, D. (2014). Perceiving crowd attention: Ensemble perception of a crowd's gaze. *Psychological Science*, *25*(10), 1903–1913.

Sweeny, T. D., Wurnitsch, N., Gopnik, A., & Whitney, D. (2015). Ensemble perception of size in 4-5-year-old children. *Developmental Science*, *18*(4), 556–568.

van Osch, Y., Blanken, I., Meijs, M. H., & van Wolferen, J. (2015). A group's physical attractiveness is greater than the average attractiveness of its members: The group attractiveness effect. *Personality and Social Psychology Bulletin*, *41*(4), 559–574.

Walker, D., & Vul, E. (2014). Hierarchical encoding makes individuals in a group seem more attractive. *Psychological Science*, *25*(1), 230–235, https://doi.org/10.1177/0956797613497969.

Wang, Y., Yao, P., & Zhou, G. (2015). The influence of facial attractiveness and personality labels on men and women's mate preference. *Acta Psychologica Sinica*, *47*(1), 108–118.

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*(8), 1293–1313.

Willenbockel, V., Sadr, J., Fiset, D., Horne, G. O., Gosselin, F., & Tanaka, J. W. (2010). Controlling low-level image properties: The SHINE toolbox. *Behavior Research Methods*, *42*(3), 671–684, https://doi.org/10.3758/BRM.42.3.671.

Zebrowitz, L. A., Hall, J. A., Murphy, N. A., & Rhodes, G. (2002). Looking smart and looking good: Facial cues to intelligence and their origins. *Personality and Social Psychology Bulletin*, *28*(2), 238–249, https://doi.org/10.1177/0146167202282009.

Zhang, W., & Luck, S. J. (2008, May 8). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233–235, https://doi.org/10.1038/nature06860.